Published in final edited form as:

Nature. 2021 October ; 598(7879): 167–173. doi:10.1038/s41586-021-03223-w.

Epigenomic Diversity of Cortical Projection Neurons in the Mouse Brain

Zhuzhu Zhang[1,*], **Jingtian Zhou**[1,2,*], **Pengcheng Tan**[1,3], **Yan Pang**[4], **Angeline C. Rivkin**[1], **Megan A. Kirchgessner**[4,5], **Elora Williams**[6], **Cheng-Ta Lee**[7], **Hanqing Liu**[1,8], **Alexis D. Franklin**[4], **Paula Assakura Miyazaki**[4], **Anna Bartlett**[1], **Andrew I. Aldridge**[1], **Minh Vu**[4], **Lara Boggeman**[9], **Conor Fitzpatrick**[9], **Joseph R. Nery**[1], **Rosa G. Castanon**[1], **Mohammad Rashid**[4], **Matthew W. Jacobs**[4], **Tony Ito-Cole**[4], **Carolyn O'Connor**[9], **António Pinto-Duartec**[10], **Bertha Dominguez**[7], **Jared B. Smith**[6], **Sheng-Yong Niu**[1], **Kuo-Fen Lee**[7], **Xin Jin**[6], **Eran A. Mukamel**[11], **M. Margarita Behrens**[10], **Joseph R. Ecker**[1,12,†], **Edward M. Callaway**[4,†]

[1]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037

[2]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA 92093

[3]School of Pharmaceutical Sciences, Tsinghua University, Beijing, China, 100084

[4]Systems Neurobiology Laboratories, The Salk Institute for Biological Studies, La Jolla, CA 92037

[5]Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA 92093

[6]Molecular Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037

[7]Peptide Biology Laboratories, The Salk Institute for Biological Studies, La Jolla, CA 92037

[8]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093

[9]Flow Cytometry Core Facility, The Salk Institute for Biological Studies, La Jolla, CA 92037

[10]Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037

[11]Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92037

[12]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037

† Correspondence and requests for materials should be addressed to E.M.C and J.R.E.: callaway@salk.edu, ecker@salk.edu.
*These authors contributed equally

Author contribution

Contribution to research design: E.M.C., Z.Z, M.M.B., J.R.E., J.Z., X.J., K.L.
Contribution to data collection: Z.Z., Y.P., A.R., E.W., C.L., M.A.K., A.F., P.A.M, A.B, A.A., M.V., L.B., C.F., J.R.N., R.G.C., M.R., M.J., T.I., B.D., J.B.S, C.O., M.M.B.
Contribution to data analysis: J.Z., Z.Z., P.T., E.M.C, M.A.K, A.F., H.L., S.N.
Contribution to data archive/infrastructure: E.A.M., Z.Z., Y.P., A.R., A.B.
Contribution to research coordination: Z.Z., E.M.C., J.R.E., M.M.B., Y.P., X.J., E.W., C.L., E.A.M., K.L.
Contribution to writing manuscript: J.Z., Z.Z., E.M.C., P.T., J.R.E., E.A.M., M.M.B.

Competing interests

J.R.E serves on the scientific advisory board of Zymo Research Inc.

Abstract

Neuronal cell types are classically defined by their molecular properties, anatomy, and functions. While recent advances in single-cell genomics have led to high-resolution molecular characterization of cell type diversity in the brain[1], neuronal cell types are often studied out of the context of their anatomical properties. To better understand the relationship between molecular and anatomical features defining cortical neurons, we combined retrograde labeling with single-nucleus DNA methylation sequencing to link neural epigenomic properties to projections. We examined 11,827 single neocortical neurons from 63 cortico-cortical (CC) and cortico-subcortical long-distance projections. Our results revealed unique epigenetic signatures of projection neurons that correspond to their laminar and regional location and projection patterns. Based on their epigenomes, intra-telencephalic (IT) cells projecting to different cortical targets could be further distinguished, and some layer 5 neurons projecting to extra-telencephalic targets (L5 ET) formed separate clusters that aligned with their axonal projections. Such separation varied between cortical areas, suggesting area-specific differences in L5 ET subtypes, which were further validated by anatomical studies. Interestingly, a population of CC projection neurons clustered with L5 ET rather than IT neurons, suggesting a population of L5 ET cortical neurons projecting to both targets (L5 ET+CC). We verified the existence of these neurons by dual retrograde labeling and by anterograde tracing of CC projection neurons, which revealed axon terminals in ET targets including thalamus, superior colliculus, and pons. These findings highlight the power of single-cell epigenomic approaches to connect the molecular properties of neurons with their anatomical and projection properties.

Main

The mammalian brain is a complex system consisting of multiple types of neurons with diverse morphology, physiology, connections, gene expression, and epigenetic modifications. Identifying brain cell types and how they interact is critical to understanding the neural mechanisms that underlie brain function. Single-cell technologies deconvolve mammalian brains into molecularly defined cell clusters corresponding to putative neuron types[1]. However, the correspondence between molecular cell types and neuronal populations defined by connectivity are largely unknown.

Previous single-cell analyses have revealed transcriptomic clusters and linked them to neuron types with different projection patterns in a few particular brain regions[2–5]. For the cerebral cortex, the most prominent molecular distinction related to projection targets is the separation of cortical neurons into distinct and apparently non-overlapping IT and L5 ET (also called pyramidal tract, PT) groups. In some cases L5 ET cells have been further divided based on both gene expression and corresponding axon projections[2]. While the separation of L5 IT and ET neurons appears to be conserved across cortical areas[6] and species[7], a systematic analysis of the relationships between a larger set of projection targets and molecular identities across multiple cortical areas has not been conducted. To what extent cortical projection neuron types can be further distinguished or divided by incorporating anatomical information with molecular analyses, and whether these cell types and correspondences are conserved across cortical areas is unclear.

Epi-Retro-Seq of 63 cortical projections

To address these questions we developed Epi-Retro-Seq, which applies single nucleus methylation sequencing (snmC-Seq)[8] to neurons dissected from cortical source regions which were labeled based on their long distance projections to specific cortical and subcortical targets (Fig. 1a). In Epi-Retro-Seq, the retrograde viral tracer rAAV2-retro-Cre[9] is injected in the target region in an INTACT mouse[10], turning on Cre-dependent nuclear-GFP expression in neurons that project to the injected target, throughout the mouse brain. Sources regions of interest are manually dissected (Methods), from which nuclei are prepared, and $GFP^+/NeuN^+$ nuclei (the GFP-labeled projection neurons) are isolated as single nuclei using fluorescence activated nuclei sorting (FANS) and assayed using snmC-Seq2[8]. SnmC-Seq has the unique ability to allow identification of potential regulatory elements and a prediction of gene expression in the same neurons[10–12]. In addition, methylation at non-CG (CH; H= A, T, C) dinucleotides (mCH) accumulates and methylation at CG dinucleotides (mCG) reconfigures during cortical synaptic development, suggesting possible links between epigenetics and connectivity[13,14].

We performed Epi-Retro-Seq to characterize projection neurons from eight mouse cortical areas (“source”) that project to ten cortical or subcortical regions (“target”), covering overall 26 CC projections and 37 cortico-subcortical projections (Supplementary Table 1). The ten injected target regions include four cortical areas [the primary motor cortex (MOp), primary somatosensory cortex (SSp), anterior cingulate area (ACA), and primary visual cortex (VISp)], and six major subcortical structures [the striatum (STR), thalamus (TH), superior colliculus (SC), ventral tegmental area and substantia nigra (VTA+SN), pons, and medulla (MY)]. The eight dissected source cortical regions are MOp, SSp, ACA, agranular insular cortex (AI), retrosplenial cortex (RSP), auditory cortex (AUDp+AUDd+AUDv), posterior parietal cortex (PTLp), and visual cortex (VISp+VISpm+VISl+VISli)] (Extended Data Fig. 1).

Methylome of cortical projection neurons

After quality control procedures (Methods), we obtained high-quality single methylomes for 11,827 cortical projection neurons (Extended Data Fig. 2). The mCH level in each single nucleus was computed across the genome using 100 kb genomic bins and used to perform unsupervised clustering of the projection neurons. Overall, the cortical projection neuron clusters were annotated into 10 subclasses (Fig. 1b) based on the reduced levels of gene body mCH, a proxy for gene expression, of known marker genes (Methods). Results from cluster analyses and annotation were used to conduct a further quality check to identify neurons whose projection targets could not be confidently assigned due to potential artifacts (Methods). We identified 1,431 neurons from experiments where the projection target could not be confidently assigned (Extended Data Fig. 2i), leaving 10,396 neurons with confident projection target assignments. All subsequent analyses that incorporate projection target information are restricted to these neurons.

Within each cell subclass, excitatory neurons but not inhibitory neurons from different cortical regions were further separated from each other (Fig. 1c), demonstrating the distinct spatial DNA methylation patterns in cortical projection neurons. The cell subclasses and

spatial patterns in Epi-Retro-Seq were in agreement with those in snmC-seq data from the same cortical regions without enrichment of specific projections (Extended Fig. 3a). Neurons projecting to different target regions were more similar within each subclass than neurons from different source regions (Fig. 1d), indicating that they shared a more similar DNA methylation landscape. Neighbor enrichment scores were used to quantify the variations of DNA methylation that originated from different cell types, cortical spatial regions, and projection targets (Methods). Neurons from the same subclass occupied highly similar regions in the dimension reduction space (neighbor enrichment score was near 1, Fig. 1e). Scores were also high for comparisons across neurons from the same source, followed by projections to the same target. Scores were near chance (neighbor enrichment score 0.5) for biological replicates, indicating that mCH profiles of different replicates are highly consistent (Fig. 1e).

Although neurons projecting to different target regions were not completely separated on t-SNE, we observed an explicit enrichment of CC and cortico-striatal projection neurons in IT subclasses (L2/3, L4, L5 IT, L6 IT, and Claustrum (CLA)), separated from neurons that project to the remaining structures outside the telencephalon which were categorized as L5 ET neurons (Fig. 1f, Extended Data Fig. 3). The enrichment is highly consistent across source regions (Extended Data Fig. 3b). As expected, many cortico-thalamic projecting neurons were also found in the L6 CT (Corticothalamic) subclass (Fig. 1f, Extended Data Fig. 3). These enrichment patterns are consistent with our knowledge about laminar enrichment of the projection neurons, which reflects the high quality of our retrogradely labeled single-nuclei methylation dataset.

To further quantify methylation differences between neurons from different source regions or projecting to different target regions, we used the area under the receiver operating characteristics curve (AUROC) of linear models trained to distinguish source pairs or target pairs based on mCH (Methods). We found that most neurons dissected from different source regions could be well separated (Fig. 1g). Most of the neurons projecting to different target regions were also separable by mCH in this supervised setting (Fig. 1g), although they were closely mixed in the unsupervised embeddings (Fig. 1d). These findings indicate that nearly all of the different types of projection neurons that were profiled have differences in their epigenomes. Further analyses of these quantitative differences, described below, allowed assessment of possible organizational principles that might exist in the relationships between DNA methylation, projections targets, and sources, including both areal and laminar sources.

Predicting IT neuron targets with mCH

In total, 42.6% of the cortical projection neurons profiled in our Epi-Retro-Seq data were identified as IT, and annotated according to their presumptive cortical layers (Fig. 1b). We investigated the contribution of the cortical area in which cell bodies were located versus their cortical projection targets, to the variation of their DNA methylation profiles. We focused on 26 CC projections from 8 cortical areas to 4 different cortical targets. All possible pairs of 4 cortical targets were assessed for each of the 8 sources to generate 32 AUROC scores, organized according to projection target pairs (Fig. 2a, Extended Data Fig. 4a–d). Among the six projection target pairs examined, neurons projecting to MOp versus

ACA were overall most distinguishable (average AUROC = 0.922), similar to neurons projecting to SSp versus VISp and ACA versus VISp (average AUROC = 0.915, 0.914), while neurons that project to SSp versus ACA and to MOp versus VISp were the least separable (average AUROC = 0.837, 0.831) (Fig. 2a). In addition, for each target pair, the performance of the predictive model varied among neurons from different source cortical regions (Fig. 2a, Extended Data Fig. 4a–d).

These analyses suggest that epigenetic differences between CC projection neurons depend on a combination of both the specific targets to which neurons project and the source region where the neurons reside. For example, among AUD IT neurons, AUD→SSp neurons were better separated from AUD→VISp neurons (AUROC = 0.974; Fig. 2b, e) than from AUD→ACA neurons (AUROC = 0.766; Fig. 2c, e). The distinctions between these projections did not stem from different distributions across layers (Fig. 2d). This demonstrates that the level of epigenetic differences between AUD IT neurons varies depending on their projection targets. On the other hand, when comparing neurons from different sources projecting to the same target pair, we observed different levels of distinguishability in our models. For example, while MOp-projecting versus ACA-projecting neurons were more distinguishable (i.e. higher AUROC scores) than SSp-projecting versus ACA-projecting neurons, we observed variation of the AUROC scores across different source regions for both target pairs (Fig. 2f, g). To further examine whether the same epigenetic differences that distinguished target pairs for one source might be conserved across sources, we trained models to predict targets using neurons from one source and then tested it on another source (Methods). Interestingly, these cross-source models can distinguish target pairs in many cases, while the performance of models trained on any particular region varied in their ability to predict projections from other regions (Fig. 2h, i, Extended Data Fig. 4e–h). For example, the model trained on AUD performed better in distinguishing VIS→MOp versus VIS→ACA neurons than the models trained on RSP or PTLp (Fig. 2h). This suggests that AUD and VIS neurons are more similar to each other in the molecular markers that distinguish neurons projecting to MOp versus ACA than other cortical areas. These results indicate that cortical regions might form different groups with shared correlations between molecular markers and projection targets.

In addition, we assessed the level of distinguishability between two cortical targets, both for neurons within the same layer and for neurons in different layers (Fig. 2j, k, Extended Data Fig. 5a–c). By training and testing the predictive models in each layer separately, we typically observed higher distinguishability between ACA-projecting versus VISp-projecting neurons than between SSp-projecting versus ACA-projecting neurons (Fig. 2j, k). But predictions for SSp-projecting versus ACA-projecting neurons were more variable, with some sources being better than others for all layers (e.g. MOp versus PTLp, Fig. 2k) and some layers being better than others, even for the same source (e.g. AUD and VIS, Fig. 2k). We further tested if cross-layer-trained models could distinguish the projection targets (Methods), and observed that the performance was generally comparable to within-layer models (Extended Data Fig. 5d–f). These results suggest that there may be shared epigenetic signatures across layers that contribute to correlations with the projection targets.

Furthermore, we identified differentially methylated genes at CH sites (CH-DMGs) between different pairs of CC projection neurons in each source region using hierarchical linear models. In total, 1,644 CH-DMGs were identified (Supplementary Table 3, examples in Extended Data Fig. 5g), among which 1,497 (91.1%) were statistically significant in only one source region. That the vast majority of CH-DMGs were unique to one source region, suggests that different genes may participate in defining projections from different source regions. Gene ontology (GO) enrichment analysis revealed that CH-DMGs were enriched for genes that participate in intracellular transport, regulation of synapse structure, etc. (Supplementary Table 3), relevant to functions that might differ for neurons with different projections. For example, Bassoon (Bsn) is differentially methylated between MOp-projecting and SSp-projecting neurons in AUD and VIS (Extended Data Fig. 5g). It encodes a presynaptic cytomatrix protein expressed primarily in neurons, and is essential in regulation of neurotransmitter release[15]. Scn2a1 encodes a voltage dependent sodium channel protein and is differentially methylated between ACA-projecting and VISp-projecting neurons in AI and PTLp (Extended Data Fig. 5g). This channel regulates neuronal excitability and variants are associated with autism and seizure disorders[16].

Epigenetically distinct L5 ET subtypes

L5 ET neurons are the most abundant cell population in our datasets (4,176 (35.3%) single neurons), and are 6.3 fold enriched in Epi-Retro-Seq compared to the total number of neurons observed in unbiased snmC-seq2 profiling. This provides us with a unique opportunity to more closely investigate subpopulations of L5 ET neurons. L5 ET neurons further segregated into 15 clusters upon uniform manifold approximation and projection (UMAP) embedding (Fig. 3a). Much of the separation between clusters was driven by the source location of the neurons, as neurons from different source regions were clearly separated on the UMAP (Fig. 3b) and each of the clusters consists of neurons mostly from one or two source regions (Extended Data Fig. 6a). The similarities between L5 ET neurons from different sources (Fig. 3c) were not well explained by their spatial proximity anterior-posteriorly or medial-laterally, but better correlated with the anatomical and functional connectivity between these regions. For example, MOp and SSp are components of the somatic sensorimotor subnetwork, while AUD, VIS, ACA, and PTLp are components of the medial subnetwork that channels information between sensory areas (that include VISp and AUD) and higher order association areas (that include PTLp and ACA)[17].

To further explore the molecular identity of these L5 ET clusters, we identified 2,675 CH-DMGs (Fig. 3d, Extended Fig. 6c, Supplementary Table 4) and 341,748 CG-DMRs (Fig. 3e, Supplementary Table 5) that were hypo-methylated in the corresponding L5 ET clusters. Gene ontology (GO) enrichment analysis revealed that these CH-DMGs were enriched in genes involved in cell communication, neurogenesis, cell morphogenesis, and axon guidance (Supplementary Table 4). The average length of CG-DMRs was 227 bp, and 84.9% of them were distal elements that located more than 5kb from the annotated transcription start sites (TSSs). The level of mCH at gene bodies is inversely correlated with gene expression, while the level of mCG at gene regulatory elements, such as promoters and enhancers, is inversely correlated with their regulatory activities. These relationships allowed us to use a gene regulatory network-based method to integrate this information and identify

transcription factors (TFs) that might function as key regulators in each cluster (Methods; Fig. 3f, Extended Data Fig. 6d, e). For example, *Rora* (RAR Related Orphan Receptor A), a transcriptional activator, was scored as one of the top TFs and is hypo-CH-methylated in clusters 1, 8, and 13, and especially in cluster 8, indicating its potential expression. The binding motif of RORA was also enriched in the CG-DMRs of these same clusters, suggesting that RORA may bind to cis-regulatory elements that in turn regulate a set of predicted downstream target genes. Many of these target genes are related to brain functions and also hypo-methylated in cluster 8 (Extended Data Fig. 6f).

L5 ET subtypes project differently

Neurons from the same source regions (except AI and RSP) distributed into more than one cluster (Fig. 3a, b, Extended Data Fig. 6b), prompting us to ask whether some of the differences between L5 ET clusters also correspond to the different projection targets. To investigate this, we performed another iteration of cluster analysis using L5 ET cell data from each of the source regions separately, and identified finer L5 ET clusters within each source region (Extended Data Fig. 7a).

Among all comparisons between projection targets and clusters, neurons projecting to medulla (MY) were most distinct. SSp L5 ET neurons further segregated into seven clusters (Fig. 4a), among which SSp→MY neurons showed a clear enrichment in cluster 0 (FDR = 3.69E-2, Wald test; Fig. 4b, c). Similarly, we identified seven clusters of MOp L5 ET neurons, and MOp→MY neurons were also significantly enriched in one of the clusters (FDR = 1.44E-2, Wald test; Extended Data Fig. 7c, d). Moreover, MY-projecting neurons were robustly distinguished from other L5 ET neurons in our prediction models for both MOp and SSp (average AUROC = 0.929, 0.864; Extended Data Fig. 8a). To investigate which genes drive the observed epigenomic differences between MY-projecting L5 ET neurons and other L5 ET neurons, we identified 1,380 (293) CH-DMGs between MOp(SSp)→MY L5 ET neurons and at least one of the other ET projections (Fig. 4d, e, Supplementary Table 6). Among these, 180 CH-DMGs were identified in both MOp→MY and SSp→MY neurons (examples highlighted in Fig. 4d, e), suggesting a general regulatory mechanism that may be shared by different cortical regions. Accordingly, models trained in either MOp or SSp to distinguish MY-projecting neurons usually performed well when tested in the other region (Extended Data Fig. 8b). Indeed, similar enrichment of MY-projecting neurons in subpopulations of L5 ET neurons has been reported in ALM using scRNA-seq (retro-seq)[6]. To compare these observations, we used gene body mCH as a proxy for gene expression to integrate our L5 ET Epi-Retro-Seq data with the ALM retro-seq data. Joint t-SNE showed that the MY-projecting L5 ET neurons were enriched in the same cluster (Extended Data Fig. 9). *Slco2a1*, a marker gene of the ALM MY-projecting cluster is hypo-methylated in MOp→MY but not in SSp→MY neurons (Extended Data Fig. 9h). We identified *Astn2* as a marker gene for the MY-projecting L5 ET cluster in both MOp and SSp (Extended Data Fig. 9i). ASTN2 mediates the recycling of neuronal cell adhesion molecule ASTN1 in migrating neurons[18], and its deletion has been associated with neurodevelopmental disorders[19]. This suggests that, compared to other L5 ET neurons, MY-projecting neurons have distinct molecular properties, and these distinctions are likely shared across several cortical regions.

In addition to the MY-projecting L5 ET neurons, we also observed differences in genome-wide mCH profiles between other ET projections. For example, L5 ET neurons in AI were segregated into five clusters (Fig. 4f), and AI→pons and AI→SC neurons were enriched in different clusters (Fig. 4g, h, Extended Data Fig. 7c, 8c). In contrast, AI→pons and AI→TH neurons were enriched in similar clusters (Extended Data Fig. 7c, 8c). Analysis of gene body mCH identified 145 CH-DMGs that were differentially methylated between AI→SC neurons versus AI→pons, while most of them had similar methylation patterns between AI→pons and AI→TH neurons (Fig. 4i, Supplementary Table 6). Together, the results suggest that AI→pons neurons are more distinct from AI→SC neurons and are similar to AI→TH neurons.

In contrast to the conservation across cortical areas ALM, MOp, and SSp for differences related to projections to MY, differences between pons-projecting and SC-projecting neurons were not conserved across all cortical areas. The prediction model trained to distinguish pons- versus SC-projecting neurons performed well in distinguishing them from cortical regions AI (AUROC = 0.939) and VIS (AUROC = 0.868), but performed poorly in PTLp neurons (AUROC = 0.726) (Extended Data Fig. 8a). The AUROC scores were correlated with the counts of CH-DMGs identified between SC-projecting versus pons-projecting neurons in the corresponding source regions (Spearman r=0.683). We further hypothesized that in a cortical area where more neurons project to both pons and SC, the epigenetic profiles of pons- and SC-projecting neurons are less distinguishable, and vice versa. To test this hypothesis, we performed double retrograde labeling of pons and SC, and counted in each cortical source region the number of neurons labeled only by the tracer injected into pons, only SC, or both (Supplementary Table 7). Indeed, PTLp had the highest percentage of double-labeled neurons, and in general the AUROC score from our model was negatively correlated with the proportion of double-labeled cells (Spearman r=−0.829, p=0.04) across the cortical regions (Extended Data Fig. 8d). These correspondences are weak, however, for most source regions, so the correlation is driven primarily by the data from PTLp.

L5 ET+CC neurons

Intriguingly, we noticed more than 30 VISp-projecting neurons in L5 ET clusters from ACA and RSP datasets (Fig. 5a, b). Since neurons in the L5 ET cluster are expected to project to ET targets, this finding suggested that some L5 neurons might project to both cortical and ET targets. These neurons were enriched specifically in one cluster in ACA and RSP, respectively (FDR = 4.88E-5, 3.34E-3, Wald test; Fig. 5a–d). This type of cluster in both RSP and ACA was marked by *Ubn2* (Extended Data Fig. 10a, top), a highly expressed gene in visual systems, and many other genes also distinguished this cluster in either region (Extended Data Fig. 10a, bottom).

Although ET cells are generally thought to lack projections to other cortical areas, there is some evidence for such cells from previous studies[20–24]. To anatomically validate our findings for RSP→VISp ET neurons in mice, we injected AAVretro-Cre in VISp and AAV-flex-GFP (Cre-dependent GFP) in RSP (Fig. 5e) or ACA in three mice (Extended Data Fig. 10b). This resulted in labeling of the complete axonal and dendritic arbors of RSP→VISp or ACA→VISp neurons such that their long-distance projections to locations

other than VISp could be assessed. For the RSP cases, we observed strong GFP labeling of axon terminals in subcortical ET regions, including TH, SC, and pons, in all three mice (Fig. 5f). For the ACA cases, axon labeling in subcortical ET regions was weaker but still readily apparent in TH (Extended Data Fig. 10b). These results indicate that single neurons in L5 of RSP and ACA can project simultaneously to both cortical and subcortical, ET targets in mice. Because these cells genetically cluster with L5 ET cells, we consider them a subtype of L5 ET cells that we refer to as L5 ET+CC. We do not use the term L5 ET+IT because many L5 ET neurons are known to project to another part of the telencephalon, the striatum.

To further assess and quantify the prevalence of L5 ET+CC cells in ACA, RSP and additional cortical areas, we made dual retrograde tracer injections into pons (CTB Alexa Fluor 647) and VISp (CTB Alexa Fluor 488) of two mice (Fig. 5g). Injections were made into topographic locations in pons known to receive input from ACA and RSP. Accordingly, overlapping retrogradely labeled neurons were observed in both ACA and RSP, allowing assessment of the proportion of double-labeled neurons within the overlap regions. Overlapping labels were also observed and quantified in higher visual cortical areas lateral and medial to VISp. A surprisingly high proportion of RSP neurons projecting to pons, 26.6%, were double-labeled (Fig. 5h, Supplementary Table 8). Substantial but smaller proportions were observed in ACA (7.0%, Fig. 5h, Supplementary Table 8) and lateral and medial higher visual areas (13.1% and 14.6%, respectively, Extended Data Fig. 10c, Supplementary Table 8).

Discussion

Here, we have quantitatively analyzed and compared the methylation of mouse cortical neurons projecting to different cortical and subcortical target regions. We identified differences between both IT neurons projecting to different cortical areas and between L5 ET neurons projecting to different ET targets. Cortical IT neurons projecting to different cortical targets were variable in the extent of their epigenetic differences. Differences between projection target pairs were typically larger than differences between cortical source areas for any given pair of projection targets. Most distinct amongst the L5 ET neurons were those projecting to the medulla. This difference has been described previously for neurons in cortical area ALM[2] and we find that this difference is conserved across the additional cortical areas that we analyzed, including MOp and SSp. In contrast, differences between L5 ET neurons projecting to SC versus pons were more distinct in some cortical areas (e.g. AI) than in others (e.g. PTLp).

We found that a subpopulation of cortico-cortical RSP→VISp and ACA→VISp neurons clustered with L5 ET cells, contrary to the expectation that L5 ET and IT cortico-cortical cells are distinct populations. This suggested that some L5 ET cells might project to cortical targets and this hypothesis was validated anatomically. Our anatomical experiments showed that RSP→VISp cells do in fact project to many ET targets, including TH, SC and pons, and we refer to this cell type as L5 ET+CC. Although we found CC projection neurons that clustered with L5 ET cells for only two of the 26 CC projections that we sampled, there remain many other combinations that we did not test. For example, our double retrograde labeling studies identified L5 ET+CC neurons in visual cortical areas lateral and medial

to VISp. Furthermore, previous studies have described L5 ET+CC cells in primary and secondary motor cortex[21,22]. It is therefore likely that future studies will reveal L5 ET+CC neurons in additional cortical areas projecting to various combinations of ET and cortical targets.

Finally, this large-scale effort linking methylation status directly to projection targets of mouse cortical neurons, allowed us to identify differences between projection cell types in TFs linked to differentially methylated regions. These observations provide insight into genetic mechanisms that might contribute to the differences in morphology and function of these cell types. As we have illustrated, this large dataset also provides the opportunity to predict regulatory elements that might be harnessed in future studies to target transgene expression to these cell types.

Methods

Experimental Animals.

All experimental procedures using live animals were approved by the Salk Institute Animal Care and Use Committee. The knock-in mouse line, R26R-CAG-loxp-stop-loxp-Sun1-sfGFP-Myc (INTACT) was used for most experiments[10] and they were maintained on a C57BL/6J background. 42–49 day old adult male and female INTACT mice were used for the retrograde labeling experiment. Adult C57BL/6J “wild-type” mice were used for double-retrograde labeling experiments.

Surgical Procedures for Viral Vector and Tracer Injections.

To label neurons projecting to regions of interest, injections of rAAV2-retro-Cre (produced by Salk Vector Core or Vigene, 2×10^{12} to 1×10^{13} viral genomes/ml, produced with capsid from Addgene plasmid #81070 packaging pAAV-EF1a-Cre from Addgene plasmid #55636) were made into both hemispheres of the INTACT mice. Animals were anesthetized with either ketamine/xylazine or isoflurane, placed in a stereotaxic frame, and 0.1 to 0.5 microliters of AAV was injected by pressure into stereotaxic coordinates corresponding to the desired projection target. A list of injection coordinates and volumes is provided in Supplementary Table 1. At least 2 male and 2 female mice were injected for each projection target. To label RSP or ACA neurons that project to VISp, VISp was injected with rAAV2-retro-Cre, and either RSP or ACA was injected with AAV-FLEX-GFP (Salk Vector Core) in each of 6 adult (3 RSP and 3 ACA), Ai14 mice. Therefore, RSP or ACA→VISp neurons, including their axonal projections, were selectively labeled with GFP. If RSP or ACA→VISp neurons also project to ET targets (L5 ET+CC neurons exist), GFP-labeled axons would be expected in subcortical ET targets such as SC, pons, and TH.

Assessment of Double-Retrograde Labeling.

To assess double-labeling of cortical cells projecting to pons and/or Superior Colliculus, or projecting to pons and/or VISp, stereotaxic pressure injections of 0.1–0.2 microliters of 0.25–0.5% of Cholera Toxin Subunit B (CTB), Alexa Fluor 488 or 647 conjugated (Molecular Probes), were successfully made into the pons and into SC of 4 mice, or into pons and VISp of 2 mice. 6–7 days later, animals were perfused with phosphate buffered

saline (PBS) followed by 4% paraformaldehyde in PBS. Brains were removed and sectioned coronally at 40 microns thickness with a freezing microtome. Sections were mounted and imaged with a 20X epifluorescence objective and images assessed to identify single and double-labeled neurons that were assigned to cortical areas. Sections with less than 5 labeled cells from either one of the injections were excluded as were sections in which there were not at least 10 labeled cells from one of the injections. Therefore, some cortical areas in which there was minimal or no overlap are not included. For each animal, double labeled cells were quantified for each region and expressed either as the proportion of double-labeled divided by the sum of all labeled cells (pons and SC) or as the proportion of double-labeled cells divided by the number of cells labeled from the pons (pons and VISp). Mean values from the 4 SC/pons animals are plotted in Fig. 4k. Values from the 2 pons/VISp animals are plotted in Fig. 5h and Extended Data Fig. 10c.

Brain dissection.

Approximately two weeks after the AAVretro injection, brains were extracted from the 56–63 day old INTACT mice, immediately submerged in ice-cold slicing buffer (2.5mM KCl, 0.5mM $CaCl_2$, 7mM $MgCl_2$, 1.25mM NaH_2PO_4, 110mM sucrose, 10mM glucose and 25mM $NaHCO_3$) that was bubbled with carbogen, and sliced into 0.6 mm coronal sections starting from the frontal pole. From each AAVretro-injected brain, the slices were kept in the ice-cold dissection buffer from which selected brain regions (Supplementary Table 1) were manually dissected under a fluorescent dissecting microscope (Olympus SZX16), following the Allen Mouse Common Coordinate Framework (CCF), Reference Atlas, Version 3 (2015) (Extended Data Fig. 1). Olympus cellSens Dimension 1.8 was used for image acquisition. The dissected brain tissues were transferred to prelabeled microcentrifuge tubes, immediately frozen in dry ice, and subsequently stored at −80°C.

Nuclei preparation and single-nucleus isolation.

For each dissected brain region, samples from 2 males and 2 females (except AI→pons - 2 male mice only)were pooled separately as biological replicates for nuclei preparation. The 2-mL glass tissue dounce homogenizer and pestles (Sigma-Aldrich D8938–1SET) were pre-chilled on ice. Nuclei were prepared using a modified protocol as reported by Lacar et al., 2016[26]. In summary, the frozen brain tissues were transferred to the dounce homogenizer with 1 mL ice-cold NIM buffer (0.25M sucrose, 25mM KCl, 5mM $MgCl_2$, 10mM Tris-HCl (pH7.4), 1mM DTT (Sigma 646563), 10µl of protease inhibitor (Sigma P8340)), with 0.1% Triton X-100 and 5µM Hoechst 33342 (Invitrogen H3570), and gently homogenized on ice with the pestle 10–15 times. The homogenate was transferred to pre-chilled microcentrifuge tubes and centrifuged at 1000 rcf for 8 min at 4°C to pellet the nuclei. The pellet was resuspended in 1 mL ice-cold NIM buffer, and again centrifuged at 1000 rcf for 8 min at 4°C. The pellet was then resuspended in 450 µL of ice-cold NSB buffer (0.25M sucrose, 5mM $MgCl_2$, 10mM Tris-HCl (pH7.4), 1mM DTT, 9ul of Protease inhibitor), and filtered through 40µM cell strainer. The filtered nuclei suspension was incubated on ice for at least 30 minutes with 50µl of nuclease-free BSA for at least 10 minutes, then incubated with GFP antibody, Alexa Fluor 488 (Invitrogen, A-21311, 1:500 dilution) and anti-NeuN antibody (EMD Millipore MAB377, 1:300 dilution) conjugated with Alexa Fluor 647 (Invitrogen A20173). $GFP^+/NeuN^+$ single nuclei were isolated using fluorescence-activated nuclei

sorting (FANS) on a BD Influx sorter with 100μm nozzle, and sorted into 384-well plates preloaded with 2μl of digestion buffer for snmC-seq2[8] (20 mL digestion buffer consists of 10 mL M-digestion buffer (2×, Zymo D5021–9), 1 ml Proteinase K (20 mg, Zymo D3001–2-20), 9 mL water, and 10 μL unmethylated lambda DNA (100 pg/μL, Promega, D1521)). The collected plates were incubated at 50°C for 20 minutes then stored at −20 °C. BD Influx Software v1.2.0.142 was used to select cell populations.

snmC-Seq2 library preparation.

Nuclei from the same projection were combined in one 384-well plate for the library preparation. We assayed approximately 384 nuclei from each projection (except the MOp→SSp projection from which 768 nuclei were assayed). The bisulfite conversion and library preparation were performed following the detailed snmC-seq2 protocol as previously described[8]. The snmC-Seq2 libraries were sequenced on Illumina Novaseq 6000 using the S4 flow cell 2 × 150 bp mode. Freedom EVOware v2.7 was used for library preparation, and Illumina MiSeq control software v3.1.0.13 and NovaSeq 6000 control software v1.6.0/ Real-Time Analysis (RTA) v3.4.4 were used for sequencing.

Reads processing and quality controls.

We used the cemba-data pipeline to generate allc files from fastq files (cemba-data.rtfd.io), as described in Luo et al[12]. Specifically, the fastq files were first demultiplexed into single cells and trimmed of Illumina adaptors and 10 bp on both sides with Cutadapt[27]. The reads were mapped to mm10 INTACT mouse genome using Bismark[28] with Bowtie2 aligner for each single end separately. The reads with MAPQ smaller than 10 were excluded. Potential PCR duplicates were removed with Picard MarkDuplicates. The reads from two ends were then merged to generate allc files using call_methylated_sites function in methylpy[29]. The global mCCC level was used to estimate the non-conversion rate of bisulfite treatment. The cells with less than 500 k non-clonal reads or non-conversion rate greater than 1% were removed from further analysis.

Methylation data processing.

For each single cell, we computed the methylated CH (*mc*) and total CH (*tc*) basecalls of all 100 kb bins across the genome and all gene bodies annotated in GENCODE vM10[30]. The autosomal bins that were covered by more than 100 basecalls in greater than 95% of cells were used for further analysis. The autosomal genes that were covered by more than 100 basecalls in greater than 80% of cells were used for further analysis.

Computing posterior methylation levels.

For each cell, we calculated the mean (m) and variance (v) of the mCH level across the 100 kb bins or genes. Then a beta distribution was fit for each cell i, where the parameters were then estimated by

$$\alpha_i = m_i\left(\frac{m_i(1-m_i)}{v_i} - 1\right)$$

$$\beta_i = (1 - m_i)(\frac{m_i(1 - m_i)}{v_i} - 1)$$

We then calculated the posterior mCH of each bin by

$$ratio_{ij} = \frac{\alpha_i + mc_{ij}}{\alpha_i + \beta_i + tc_{ij}}$$

We normalized this rate by the cell's global mean methylation by

$$global_i = \frac{\alpha_i}{\alpha_i + \beta_i}$$

$$M_{ij} = \frac{ratio_{ij}}{global_i}$$

The values greater than 10 in M were set to 10. After normalization, M_{ij} is close to 1 when tc_{ij} is close to 0.

Identification of highly variable bins.

Highly variable methylation features were selected based on a modified version of the highly_variable_genes function in Scanpy[31]. In brief, since both the mean methylation level and the mean coverage of a feature (100 kb bin or gene) can impact methylation level dispersion[12], we grouped features that fall into a combined bin of mean and coverage, and then normalized the dispersion within each group. After dispersion normalization, we selected the top 2,000 features based on normalized dispersion for dimension reduction.

Removing potential doublets.

By plotting all cells on t-SNE, we noticed a cell population that was located in the center of the plot and has a greater number of non-clonal reads than the others. To remove these potential doublets, we modified scrublet[32] to adopt it to methylation data. Specifically, we first simulate the doublet cells by randomly selecting two cells in our dataset and sum the methylation/total basecalls of the two cells. Then the methylation levels of the simulated cells were computed using the posterior computing method. We simulated twice the number of doublets as the number of real cells. The top 2,000 highly variable features were selected for dimension reduction with principal component analysis (PCA) and the top 50 PCs were used to train a k-nearest neighbor (kNN) classifier (k=50) to predict a doublet score for each cell. Based on the histogram of doublet scores of real and simulated doublet cells, the cells with doublet score higher than 0.1 were removed from further analysis. After removing the potential doublets, 13,414 cells were kept for further analysis.

Cell clustering and annotation.

After removing potential doublets, the top 2,000 highly variable features were selected for dimension reduction with PCA. The top 50 PCs were used for t-SNE visualization and construction of kNN graph (G) with Euclidean distance (k=25). We use A to represent the connectivity of G, where A_{ij} is 1 if node j is among the 25 nearest neighbors of node i, otherwise 0. The edge weights of G were assigned as the jaccard distance of the connectivity matrix A. We ran Louvain clustering (https://github.com/taynaud/python-louvain) with resolution 1.2 to partition the cells into 31 clusters and merged these clusters into major cell subclasses based on known marker genes. Specifically, *Cux2*+ *Rorb*- (hypo-methylation in *Cux2* gene body and hyper-methylation in *Rorb* gene body) was annotated as L2/3; *Cux2*+ *Rorb*+ was annotated as L4; *Cux2*- *Rorb*+ and *Deptor*+ were annotated as L5 IT; *Sulf1*+ and *Sulf2*+ *Deptor*- were annotated as L6 IT; *Vat1l*+ was annotated as L5 ET; *Foxp2*+ was annotated as L6 CT; *Tle4*+ *Foxp2*- was annotated as L6b; *Tshz2*+ was annotated as NP; *B3gat2*+ was annotated as CLA; *Slc6a1*+ was annotated as Inh. The clusters with low global mCH level were annotated as non-neural cells, which were further confirmed by hyper-methylation of *Mef2c*. The 11,827 cells within neuronal cell clusters were selected for further analysis.

Inclusion criteria for confident target assignment

We implemented criteria to identify experiments in which artifacts could lead to inclusion of neurons that did not actually project to the intended AAVretro injection site. Neurons failing these criteria were excluded from analyses requiring identification of projection targets but were included for analyses related to neuron sources. Close inspection of the distribution of cells sampled from each projection across subclasses revealed two types of artifact: 1) for some weak projections very few neurons were retrogradely labeled, resulting in small proportions passing FANS gating criteria and subsequent inclusion of high proportions of cells accepted from the edges of FANS gates (“gating artifact”); 2) AAV-retro injection pipettes targeting deep structures (e.g. thalamus) passed through overlying cortical areas and directly labeled neurons rather than being taken up retrogradely from the intended target. This second artifact is apparent in previously published retro-seq data in which VISp IT neurons are prominent in putative cortico-tectal and cortico-pontine projection neuron populations (Fig. 3 and Extended Data Fig. 10 in Tasic et al. 2018[6]). This suggests that injections passed through VISp, which directly overlies pons and tectum. In our experiments, injections to SC and pons took oblique trajectories to minimize involvement of overlying cortical areas, but this was not possible for injections to VTA or TH.

Because FANS errors would be manifested in separate sorting runs, we assessed each FANS sorting case separately. To identify cases with high proportions of contaminating neurons (likely projecting to a different target than intended), for each FANS run, we counted the numbers of neurons that were observed in known on-target subclasses (O_{on}) and off-target subclasses (O_{off}). Assuming that the proportions of contaminated cells in each subclass would be similar to a sample without projection-type enrichment, we compared the observed counts to the counts from unbiased cortical samples[33] (E_{on} and E_{off}) collected from the slices in Extended Data Fig. 1. The fold-enrichment was computed as $\frac{O_{on}E_{off}}{O_{off}E_{on}}$. A one-sided

exact binomial test of goodness-of-fit was used to determine whether the enrichment of on-target cells was significant. Specifically, the P value was computed as $Pr(X \geq O_{on}; n, p)$, where $X \sim Binomial(n, p)$, $n = O_{on} + O_{off}$, $p = \frac{E_{on}}{E_{on} + E_{off}}$. Neurons from cases where the fold-enrichments were smaller than a threshold (see below) or the tests were not significant, were categorized as having unknown projection targets. The expected values are different for ET targets than for IT (including striatum) targets, so the thresholds depend on the target regions.

For each ET target, we considered L5 ET as on-target subclass and IT+inhibitory neurons as off-target. The thresholds for fold-enrichment and FDR (Benjamini-Hochberg procedure) were 5 and 0.01, respectively. This eliminated 7 out of 101 ET target sorts (285 out of 5,364 cells). For IT targets, we considered IT as on-target subclasses and L6 CT+inhibitory neurons as off-target. The thresholds for fold-enrichment and FDR (Benjamini-Hochberg procedure) were 3 and 0.05, respectively. This eliminated 30 out of 115 sorting cases (1,146 out of 6,463 neurons).

Note that these exclusion criterions are based on a simplified expectation of on target cell types, and the accuracy might be variable depending on the targets. For instance, when considering the striatum-projecting neurons, considering L6 CT as off-target might overestimate the off-target cells and make the exclusion more stringent. In addition, since the filter was applied at FANS run-level, there could also be a small percentage of off-target cells from the included runs. This should be noticed when using these dataset. We included the cell-type proportion of all projections in Extended Data Fig. 3c to help evaluate this potential noise.

Neighbor enrichment score.

The score was used to quantify the enrichment of cells that belong to the same category among the neighbors of each cell. A higher score represents the cells are more likely to form clusters with the cells belonging to the same category rather than in the other categories. The advantage of this score is that it only considers the local effect so that would remain high if the cells in a category form several different clusters that dissimilar with each other. The score was computed as follows. Euclidean distances between each pair of cells were computed using the first 50 PCs. For each cell, we found its 25 nearest neighbors in the same category, and $25r$ nearest neighbors from other categories, where r is the ratio between total number of cells in other categories and total number of cells in the same category. The area under the receiver operating characteristic (AUROC) using distances between the cell and these neighbor cells for distinguishing the categories were defined as the neighbor enrichment score of this cell. The methylation pattern of male and female mice are highly similar on autosome; therefore, the two genders were treated as replicates in the analyses. When computing the score for targets, neurons whose targets were not confidently assigned were excluded. When computing the score for replicates, AI→Pons projection which only has one replicate was excluded.

Pairwise prediction of the source and target regions.

Based on the sources, and targets, the neurons could be separated into groups. Each group contains the neurons projecting from a specific source to a specific target. To test the similarity of two groups of cells based on DNA methylation, we trained logistic regression models to predict the group label of each cell. The posterior of 100 kb-bin or gene body mCH were used as features. We used two methods to split the cells into training and testing sets, one uses random selection of half of the cells for training and the other half for testing (computational replicates), the other is based on the gender of the mice the cells were collected from (biological replicates). All results in the main figures were computed using the computational replicates, while the results using biological replicates are also provided in Extended Data Fig. 4 and 5. The results of corresponding comparisons were very similar between these two replicate-splitting methods. The area under the receiver operating characteristic (AUROC) from cross-validation was used to measure the performance of the model. The higher AUROC represents better ability of the model to present the group label, which indicated the two groups had larger mCH differences and were more distinguishable. Sci-kit learn was used for model implementation.

When the groups being studied contained cells from different subclasses (e.g. cortical projecting neurons in one source), we up-sampled the training set to make it better capture the group differences rather than the differences of cell distributions across subclasses. For example, when comparing neurons projecting to two different cortical targets, the subclass composition differences could make the model over-weight the features marking different subclasses. To get rid of this bias, we randomly repeated the neurons from the under-representing group and ensured the two groups had the sample number of training samples in each subclass. The models were then trained and tested in the same setting as mentioned above.

Several reasons could contribute to a low prediction performance. Biologically reasons would include: 1) Some neurons make projections to several targets simultaneously. These could result in the neurons being captured by multiple retrograde labeling experiments of different targets. It would be impossible to predict a single label with our pairwise models for this type of neuron. 2) Some neurons project to different target regions but have tiny epigenetic differences. To systematically distinguish 1) to 2), other anatomic and genetic validation are still needed.

Technical reasons would include: 3) The epigenetic differences between neurons projecting to different targets varies across replicates. 4) The contamination levels of some projections are relatively high, which make larger noise and hinder the models to capture real signals. 5) The sample sizes of some projections are small, which make the learning more challenging. 6) The models are not powerful enough to capture the complex differences between projections.

In this study, male and female mice were treated as biological replicates after removing sex chromosomes. Although methylation patterns of autosomes are similar, differences between genders or individuals might still exist. The small differences of performances between data splitting methods (based on computation or biological replicates) might

suggest a less notable effect contributed by 3) in those samples. If the cross source/layer predictions (described below) performed better than the within source/layer models, we would suspect that shared differences between neurons projecting to different targets exist across sources/layers, and the major reason for lower accuracies of within source/layer models might be 4) or 5). Elimination of contaminated FANS runs decreases the potential influence by 4), although there are still contaminated cells included in the dataset. To evaluate the potential limitation of 6), more carefully curated models, and accordingly more samples, would be required. Thus, given all these factors, we are generally more confident in the distinguishable target pairs when training and testing sets were split based on both computational and biological replicates. The interpretation of comparisons without biological replicates and the indistinguishable pairs would need to be more careful and are not involved in the major conclusions in this manuscript. Our study aims to provide a general view across multiple sources and targets. More detailed understanding of specific projections would require larger scale profiles on those specific projection types.

Cross source prediction.

The logistic regression models were trained to predict the projection targets in one source and tested in the other source. The training set and testing set came from either the biological or computational replicates. When using biological replicates, the final AUROC were the average of AUROCs by training in male mice in one source and testing in female mice in another source, and by training in female mice in the first source and testing in male mice in the second source. For cortical targets, we up-sampled the training set in the same way as the above section.

Note that when the models were training only in one source, they would not necessarily capture the shared features across sources to distinguish neurons projecting differently even if some shared differential features exist. However, when more differential features are shared across sources, the models are more likely to select the shared ones. Thus, the low performance in the analysis might indicate that there are less differential features shared across sources and the models majorly selected the differential features specific to one source but not another source, rather than representing none of the differential features are the same between the two sources. On the contrary, the high performances usually indicate that more differential features are shared between sources. Similar interpretation applies to the cross layer prediction in the next section.

Cross layer prediction.

This analysis was specifically for CC projection neurons to study whether the mCH differences between projection neurons were shared or distinct across layers. The logistic regression models were trained to predict the projection targets in all but one layer and tested in the one layer left out during training. The training set and testing set came from mice of different genders.

Identification of differentially CH-methylated genes (CH-DMGs).

Wilcoxon rank-sum test and t test were widely used to identify differential genes in single-cell studies[31], which consider each cell as an independent sample. However, the cells

from the same replicate, individual, or batch would be more similar than the cells from different ones. Therefore, considering all cells as independent samples would overestimate the statistical power in single-cell data. To address this problem and take the replicate-level variation into consideration, we used a linear mixed model for the differential analysis and performed paired-wise comparisons between groups. The posterior mCH level of 12,261 autosomal genes after coverage filters were used for these analyses. The posterior gene-body mCH was used as dependent variables. Each individual mouse was considered as a random effect. The global mCH levels and the gender of the mice were considered as fixed effects. Other fixed effects were determined based on the comparison. Specifically,

For DMGs between L5 ET clusters:

Gene_mCH ~ cluster + gender + global_mCH + (1 | mouse)

For DMGs between cortical targets in each source:

Gene_mCH ~ target + cluster + gender + global_mCH + (1 | mouse)

For DMGs between ET targets in each source:

Gene_mCH ~ target + gender + global_mCH + (1 | mouse)

Each gene was tested separately, and a two-sided Wald test was performed to estimate the *P* value for the effect being tested. FDR was computed for each pair of groups with the Benjamini/Hochberg process. The fold-change of each gene was computed by the average mCH across cells in one group divided by the average mCH across cells in the other group, with pseudo-counts of 0.1. The criterions for significance when testing difference variables were distinct and shown as follows. For DMGs between L5 ET clusters: absolute log fold-change greater than log1.5 and FDR smaller than 0.01. For DMGs between IT targets or between ET targets in each source: absolute log fold-change greater than log 1.25 and FDR smaller than 0.01.

Gene ontology enrichment analysis.

Gene ontology enrichment analysis was performed using the web server at http://geneontology.org/. The 12,261 genes that passed the coverage threshold mentioned above were used as background, and binomial tests were used to select the significant biological processes related to each DMG list. Note that gene ontology names are nomenclature that summarize many complex relationships between genes and their function, so we do not expect that these analyses can be used to directly infer how a particular gene contributes to neuronal function in a specific context.

Identification of differentially CG-methylated regions (CG-DMRs).

To identify DMRs, we merged the allc files of individual cells assigned to the same cluster to create a pseudo-bulk allc table for each cluster. Then we selected all the CG sites and combined the methylation on two DNA strands for each CpG site. We run methylpy[29] DMRfind to identify the DMRs and require the DMRs to contain at least 2 differentially methylated CpG sites (DMS).

Inference of crucial transcription factors (TF) with PageRank.

The method was modified from Taiji[34] to integrate the information of both gene body and regulatory regions. The 537 motifs in JASPAR 2018 non-redundant core vertebrate database[35] were used for these analyses. We scanned each of the motifs against the mm10 INTACT mouse genome with fimo[36] and *P* value cutoff as 1e-5. The DMRs between clusters were expanded 100 bp on both sides, and the ones overlapping with motifs were assigned to the corresponding TF. The DMRs were also assigned to the potential genes they regulated using GREAT[37]. The TFs were then linked with the target genes based on these DMRs that links to both the upstream TFs and the downstream genes. A gene regulation network was constructed where the nodes represented the genes and edges represented the links between TF genes and target genes.

To assign weights to the edges and initiate the node importance, the normalized $n_{cluster} \times n_{gene}$ methylation matrix (M) were min-max normalized across clusters to 0–1 by

$$N_{ij} = \frac{M_{ij} - min_{0 < j' \leq n_{gene}} M_{ij'}}{max_{0 < j' \leq n_{gene}} M_{ij'} - min_{0 < j' \leq n_{gene}} M_{ij'}}$$

, and $1 - N_i$ were used as the predicted expression of each gene in cluster i. The predicted expressions of all genes were used as starting importance I_0. Then we used a $n_{gene} \times n_{gene}$ matrix A to represent the adjacency matrix of TF-gene regulation network, where A_{ij} was assigned as the predicted expression level of gene i if gene i is a TF. To ensure an undirected propagation, we used $B = A + A^T$ as the final adjacency matrix. B was normalized by row into the transition matrix P by

$$P_{ij} = \frac{B_{ij}}{\sum_{j'=1}^{n_{gene}} B_{ij'}}$$

Next we performed a diffusion step of the PageRank scores through the network. For iteration t, the PageRank scores were computed by

$$I_t = P \times I_{t-1} + rp \times I_0$$

, where rp represents a restart probability to balance the global and local effect of the propagation on the network. The diffusion step was stopped when $|I_t - I_t| < 10^{-5}$.

Clustering of L5 ET cells in each source region.

L5 ET neurons from Epi-Retro-Seq and unbiased snmC-Seq were combined in this analysis. After the same process as clustering all cells to derive posterior mCH level and select highly variable features, the first 30 PCs were used for computing kNN (k=15) and Louvain clustering. The resolutions used for source regions were 1.6 for MOp, AI, AUD, and RSP; 2.0 for SSp and PTLp; 1.0 for VISp; and 2.5 for ACA. The resolutions were determined based on visually examining the cluster numbers and projection enrichment.

To confirm that there were epigenetic features distinguishing the clusters, we computed the differentially methylated 100 kb bins (DMBs) across all pairs of clusters using two-sided Wilcoxon rank-sum tests. The bins were defined as differential if the absolute log fold-change between clusters were greater than log 1.5, and FDR of the test smaller than 0.01. We also used AUROC>0.85 and AUPR>0.6 to define DMBs, which provided similar results. Two clusters in RSP that had less than 5 DMBs were merged.

Tests of projection enrichment in clusters.

As described above, the cells from the same replicate would be more similar, and considering all cells as independent samples will overestimate the statistical power in single-cell data. Therefore, we used linear mixed models to test for significant enrichment of particular projections in each cluster, considering the mouse where the cells came from. The subclsuter was used as dependent variables. Each individual mouse was considered as a random effect. The projection target was considered as fixed effects. [Cluster ~ Target + (1 | mouse)]

Each projection target and each cluster were tested separately, and two-sided Wald tests were performed to estimate the P value for the effect being tested. FDR was computed for each source with the Benjamini/Hochberg process. (Obs-Exp)/Exp in the enrichment matrices were computed using the same method as in Pearson's chi-square test.

Integration of Epi-Retro-Seq and Retro-Seq.

Single-cell transcriptomic data from Tasic 2018[2,6] was downloaded from NCBI Gene Expression Omnibus (GSE115746). 365 cells within clusters of 'L5 PT ALM *Npsr1*', 'L5 PT ALM *Slco2a1*', and 'L5 PT ALM *Hpgd*' were selected for integration analysis. The raw data was preprocessed using Scanpy[31]. Specifically, the read counts were normalized by the total read counts per cell and log transformed. Top 10,000 highly variable genes were identified and z-score scaled across all the cells. For methylation data, the posterior methylation levels of 12,261 genes in the 4,176 L5 ET cells were z-score scaled across all the cells and used for integration. We used Scanorama[38] to integrate the z-scored expression matrix and minus z-scored methylation matrix with sigma equal to 100.

Overlap score.

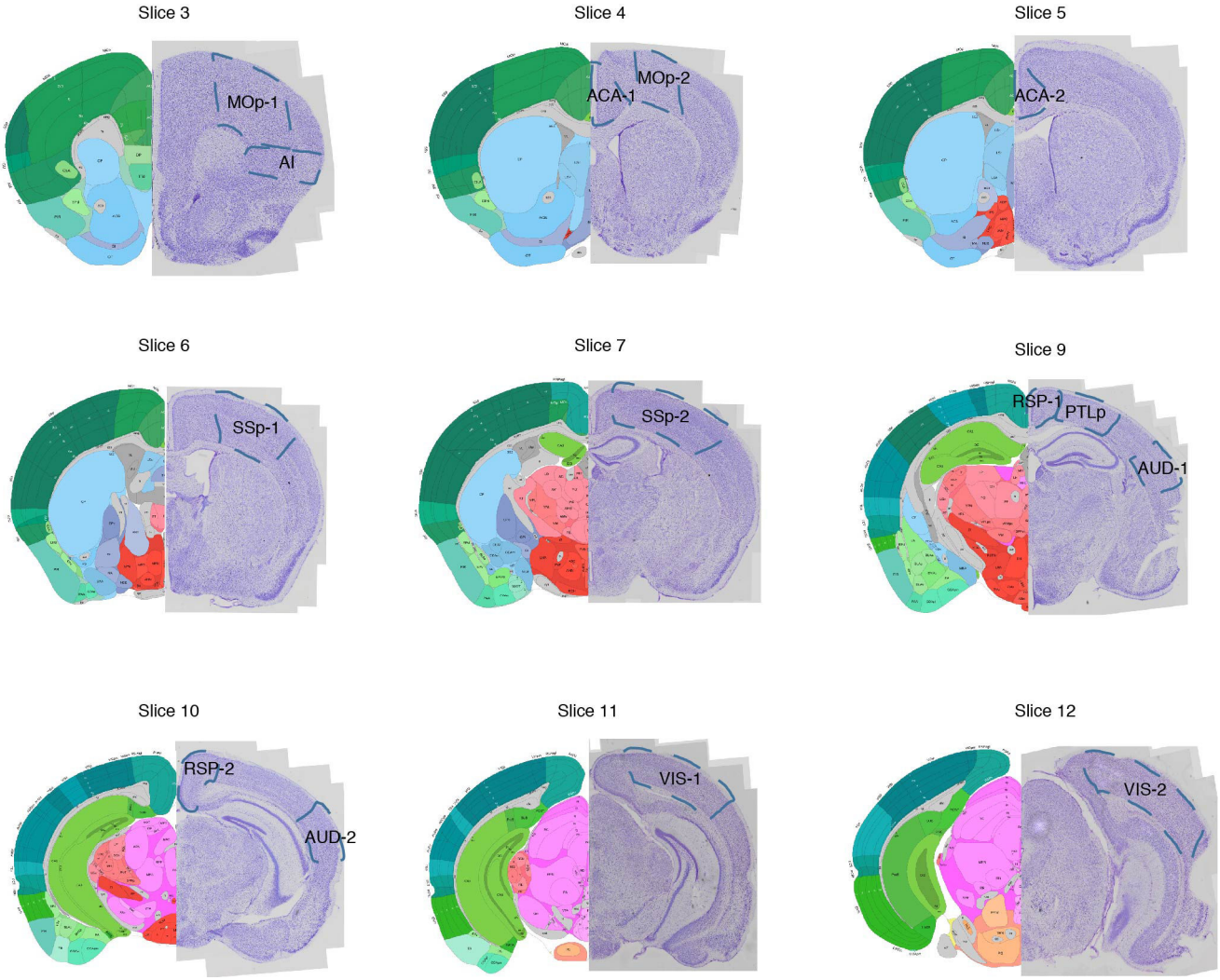
Overlap score quantifies the similarity of the distributions of two groups of cells across clusters, where higher scores represent the two groups are more likely to be co-clustered. The scores were computed using the same method as in Hodge et al[7]. Specifically, a $n_{group} \times n_{cluster}$ matrix C was first computed, where C_{ik} represents the number of group i cells in cluster k. C was normalized by row to D, and the overlap score between group i and group j was defined as $\sum_{k=1}^{n_{cluster}} min(D_{ik}, D_{jk})$.

Data access and code availability

Single cell raw and processed data included in this study were deposited to NCBI GEO/SRA with accession number GSE150170 and the NeMO ftp archive: http://data.nemoarchive.org/biccn/lab/callaway/projection/sncell/. The code for all of the analyses can be found at

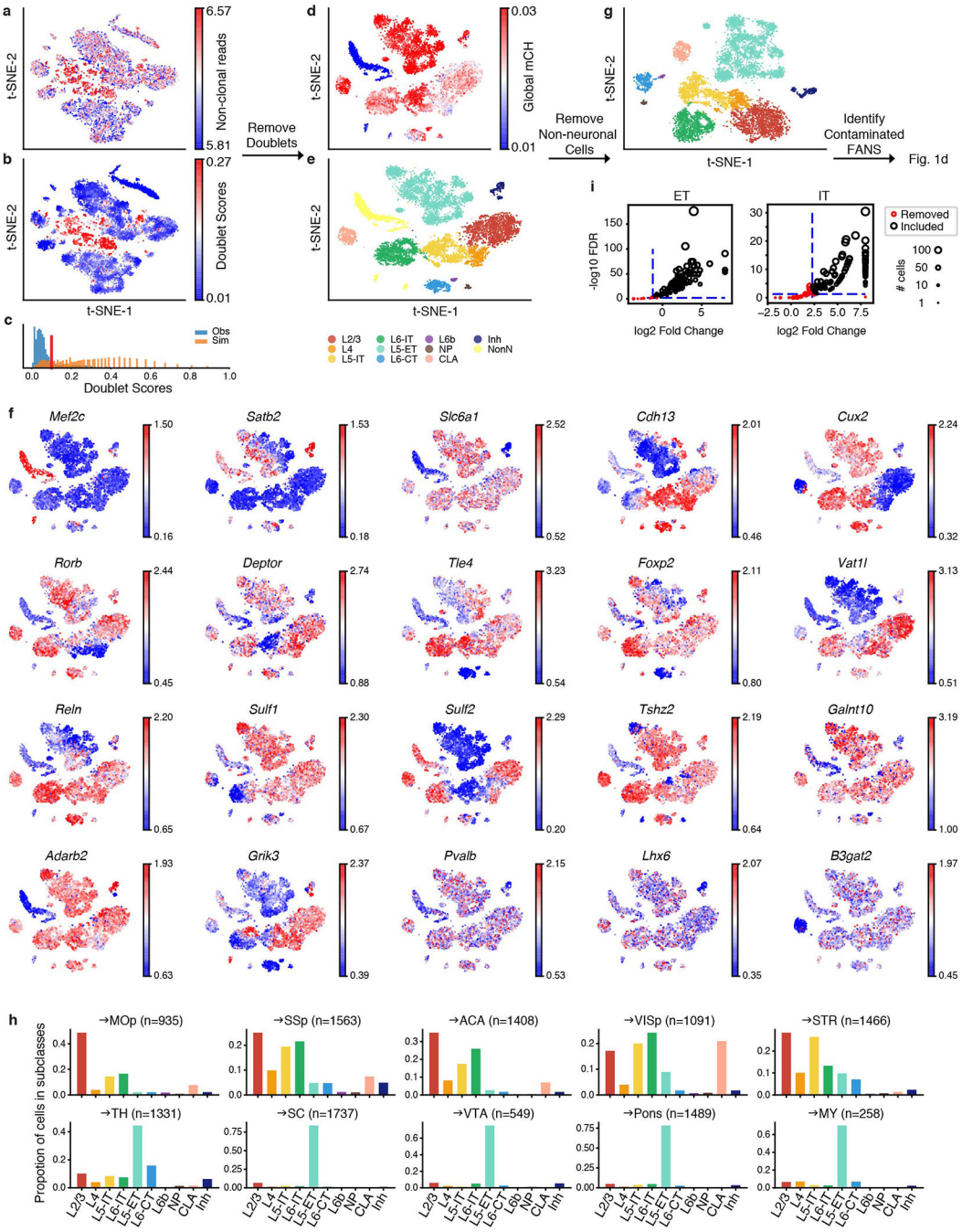
https://github.com/zhoujt1994/EpiRetroSeq2020.git. Another datasets used in this study includes the JASPAR motif database (http://meme-suite.org/db/motifs) and retro-seq data from GSE115746.

Extended Data



Extended Data Fig. 1. Source region dissection maps.

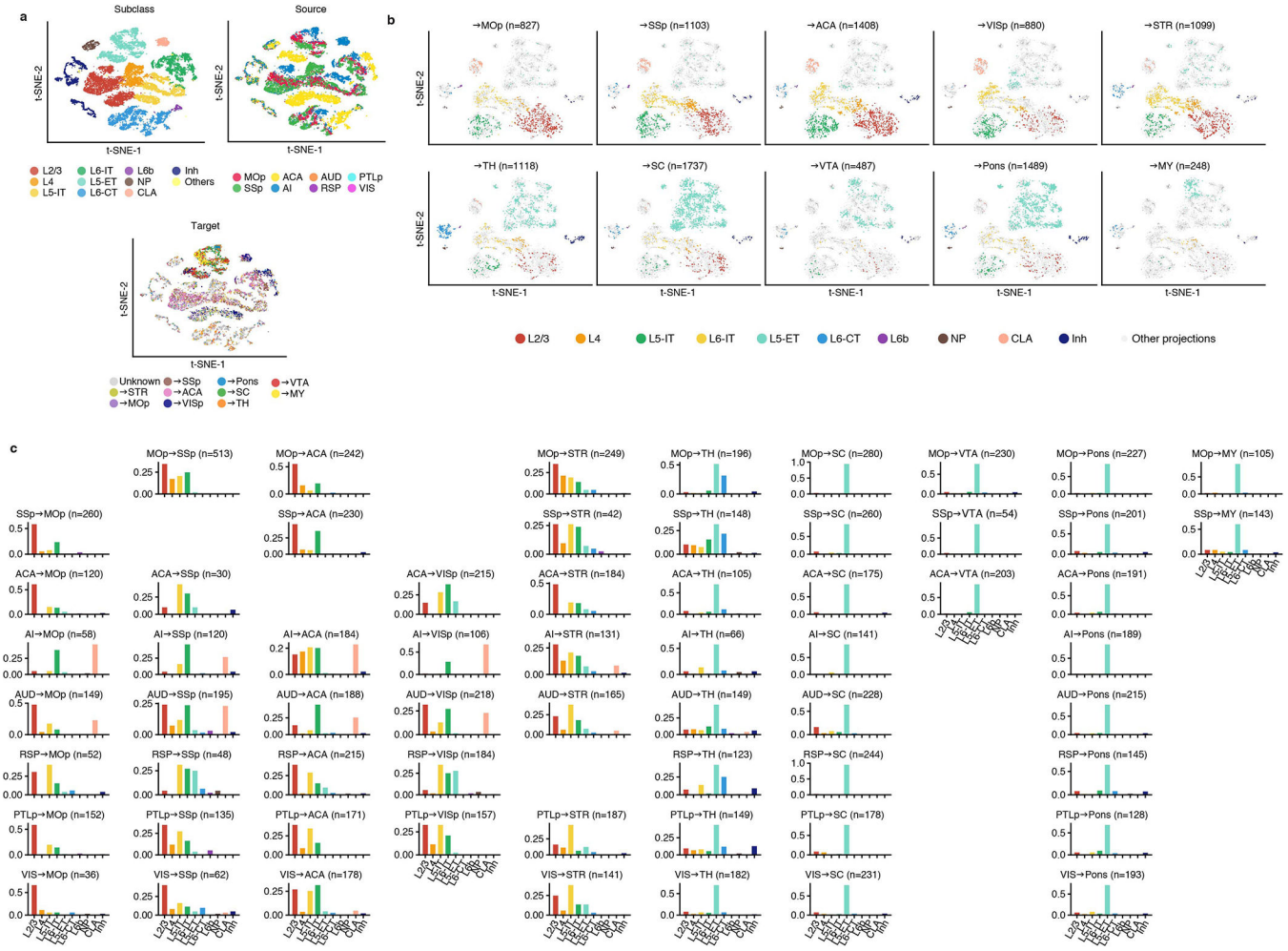
The posterior views of dissected slices are shown. The slices correspond to Allen Reference Atlas level 33~39 (slice 3), 39~45 (slice 4), 45~51 (slice 5), 51~57 (slice 6), 57~63 (slice 7), 69~75 (slice 9), 75~81 (slice 10), 81~87 (slice 11), and 87~93 (slice 12), respectively. All brain atlas images were created based on © 2017 Allen Institute for Brain Science. Allen Brain Reference Atlas. Available from: atlas.brain-map.org and Wang et al.[25]



Extended Data Fig. 2. Removing potential doublets and non-neuronal cells.

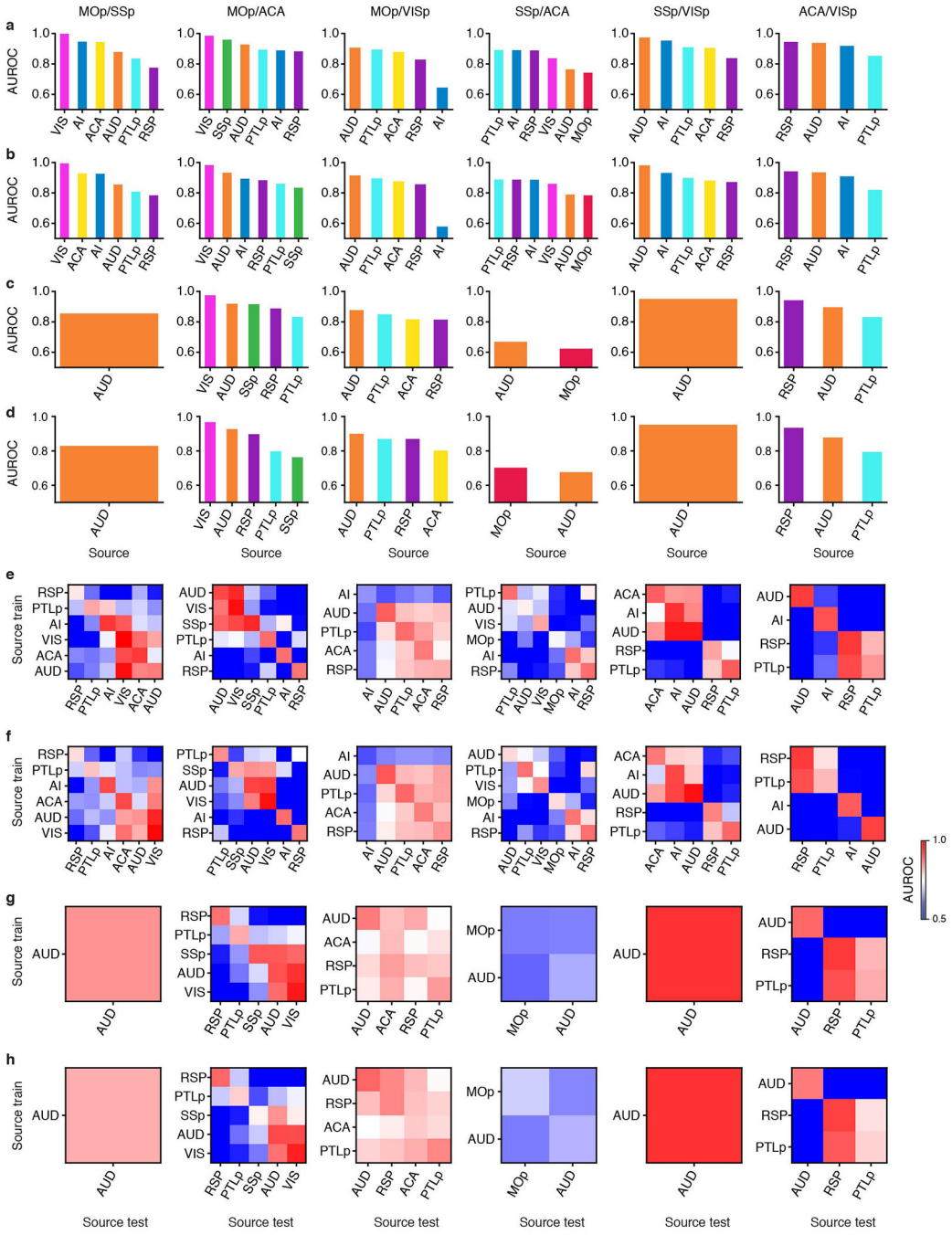
a, b, t-SNE of cells after quality control (n=16,971) colored by number of non-clonal reads (**a**) and predicted doublet scores (**b**). **c,** Distribution of doublet scores for real cells (blue) and simulated doublets (orange). **d-f,** t-SNE of cells after removing doublets (n=13,414) colored by global mCH (**d**), subclass (**e**), or normalized gene-body mCH level of known cell type gene markers (**f**). Cells with low global mCH level are usually non-neuronal cells. **g,** t-SNE of single neurons (n=11,827) colored by subclass. **h,** Proportion of single neurons in each subclass for each projection. **i,** The scatter plots for filtering FANS runs with high

contamination. Each dot represents a single run (n=101 left, 115 right), and the size of dot represents the number of on-target cells selected by the run.



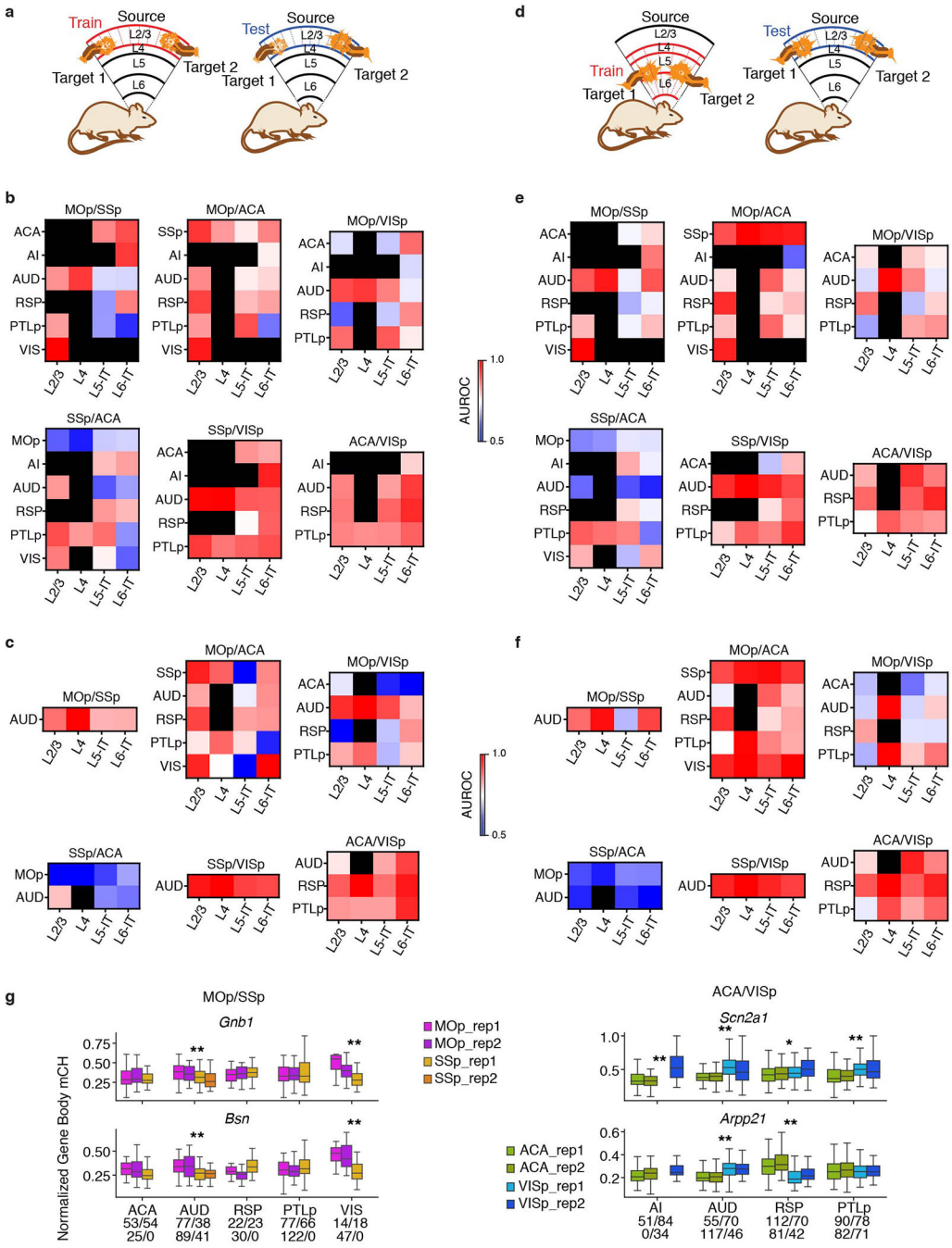
Extended Data Fig. 3. Cell type composition of all projections.

a, Joint t-SNE of neurons profiled by Epi-Retro-Seq (n=6,362) and unbiased snmC-seq2 (n=15,782, without enrichment of projections) from MOp, SSp, ACA and AI, colored by subclass (top left), source region (top right), and projection targets in Epi-Retro-Seq (bottom). **b,** t-SNE of neurons (n=11,827) projecting to each IT target (top) and ET target (bottom). The cells projecting to the target were colored by subclass and cells that project to all other targets or whose target was not confidently assigned were greyed. **c,** The proportion of cells projecting from each source (row) to each target (column) in all subclasses.



Extended Data Fig. 4. AUROC of cortical target pairs within and cross source regions.

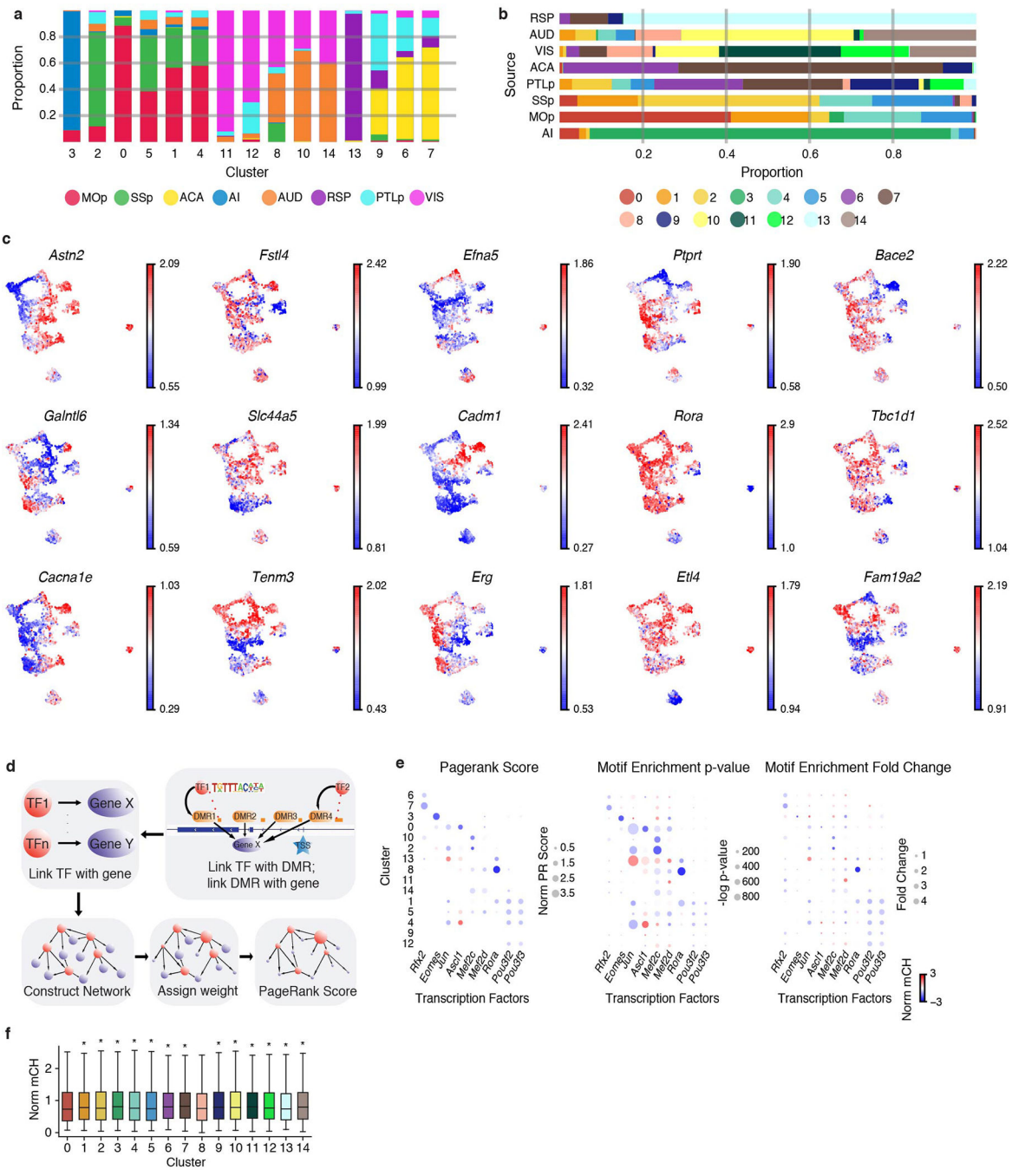
AUROC of models trained and tested in the same source region (**a-d**) or models tested in all source regions after trained in each one of them (**e-h**). Gene body (**a, c, e, g**) or 100 kb bin (**b, d, f, h**) mCH was used as features. The training and testing sets were randomly split (**a, b, e, f**) or split based on biological replicates (**c, d, g, h**). The values in (**a-d**) correspond to the diagonals of (**e-h**) but ordered decreasingly.



Extended Data Fig. 5. AUROC of cortical target pairs within and cross layers.

Demonstration of training and testing data for within layer prediction (**a**) and cross layer prediction (**d**). In (**a**), the models were trained and tested in the same layer with different cells. In (**d**), the testing sets were the same as (**a**), but the models were trained in all other layers. AUROC of within layer prediction (**b, c**) or cross layer prediction (**e, f**). The training and testing sets were randomly split (**b, e**) or split based on biological replicates (**c, f**). Gene level mCH were used for all the predictions. **g,** Boxplots of example genes that were differentially methylated at CH sites (CH-DMGs) between →MOp versus →SSp neurons

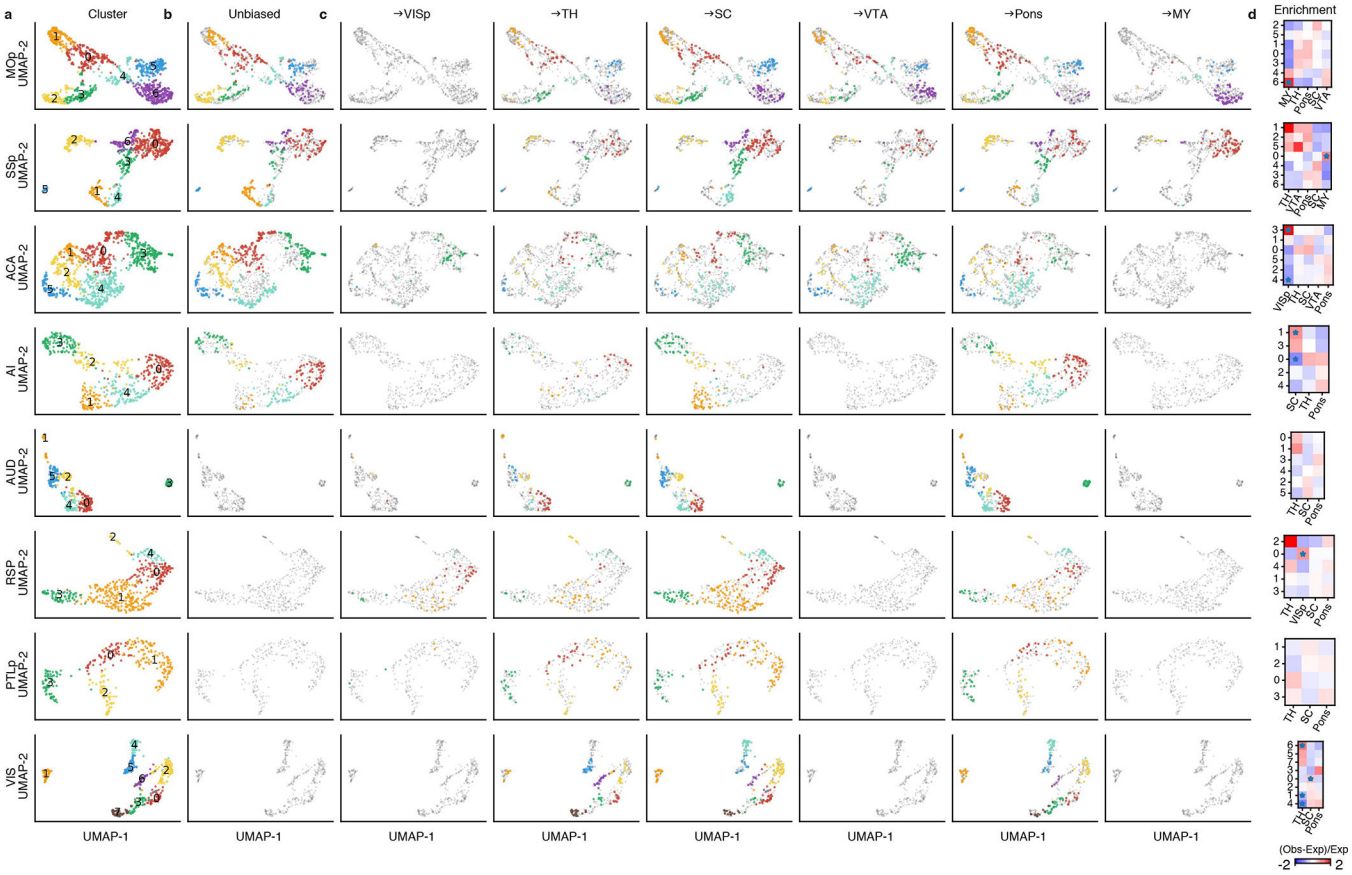
(top), or between →SSp versus →VISp neurons (bottom). The sample sizes are shown as ticklabels of x-axis. ** represents false discovery rate (FDR)<0.01 and * represents FDR<0.1. The elements of all box-plots are defined as: center line, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range.



Extended Data Fig. 6. Signature genes and TFs of L5 ET clusters.

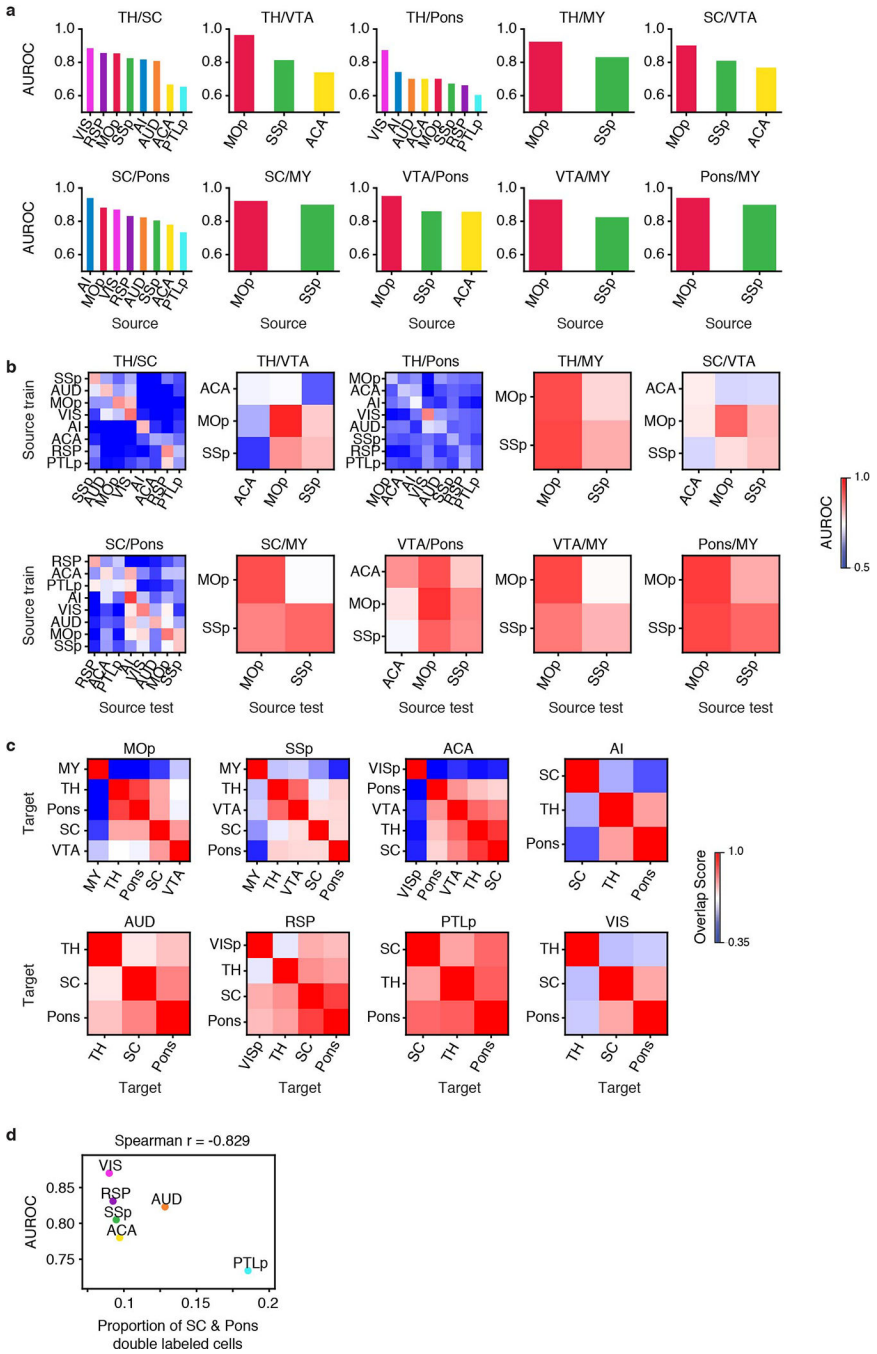
a, Proportion of cells from all source regions in each cluster. **b,** Proportion of cells in all clusters from each source region. **c,** t-SNE of L5 ET cells (n=4,176) colored by the

normalized gene-body mCH level of cluster gene markers. **d,** Workflow of the PageRank algorithm to infer crucial transcription factors. **e,** Gene body mCH (color) against PageRank score (size, left), motif enrichment *P* value (size, middle), and motif enrichment fold-change (size, right) for the example TFs in all L5 ET clusters. *P* values were computed by Homer using one-sided binomial tests. **f,** Gene body mCH in all clusters of *Rora* target genes identified in cluster 8 (n=3,299). Significances were determined by comparing cluster 8 with each of the other clusters (two-sided Wilcoxon signed-rank test, Benjamini-Hochberg FDR). * represents FDR<1e-2. The elements of all box-plots are defined as: center line, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range. FDR for all boxes are 0.60, 1.95e-25, 3.56e-12, 5.24e-29, 1.57e-10, 8.44e-09, 2.94e-32, 3.56e-41, 1.0, 1.16e-35, 5.85e-29, 2.28e-42, 1.47e-28, 6.42e-03, 1.50e-26.



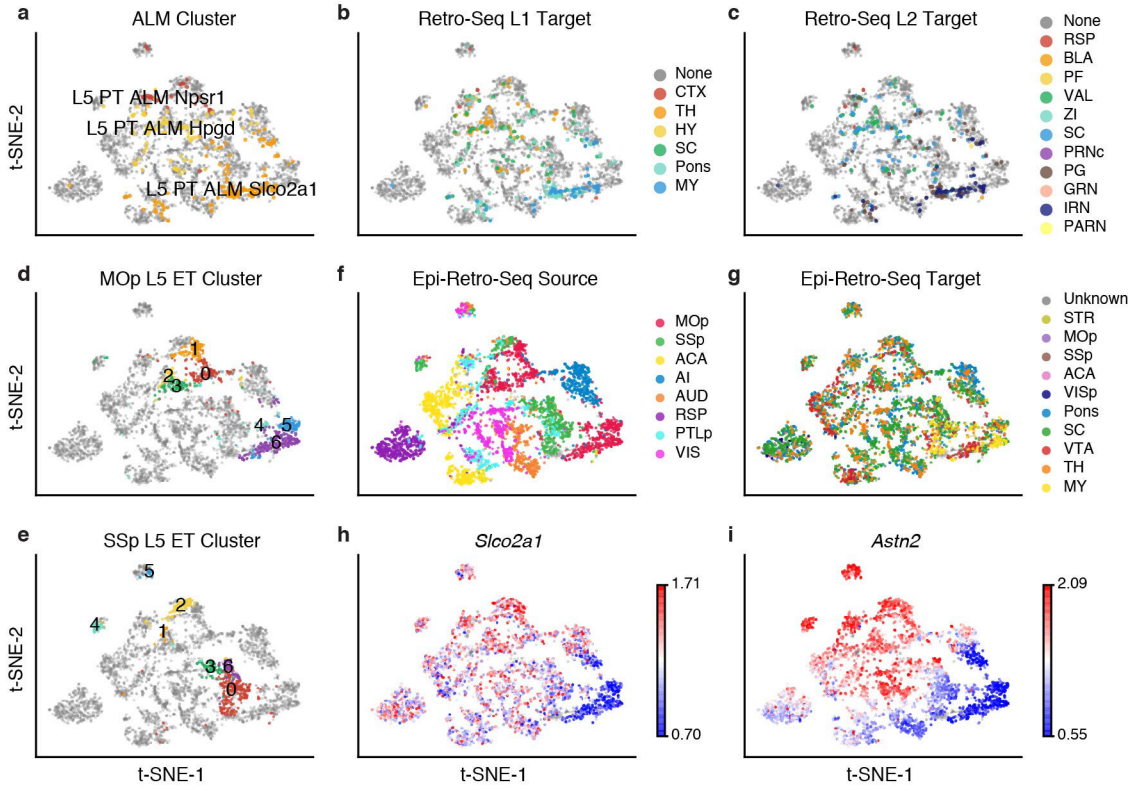
Extended Data Fig. 7. Enrichment of different projections in L5 ET clusters.

a-c, t-SNE of L5 ET cells from each source region colored by clusters. The colored cells are all cells (**a**), unbiased snmC-Seq cells (**b**), and cells projecting to each target (**c**). Other cells were greyed. **d,** The enrichment of each projection in each L5 ET cluster in each source. * represents FDR<0.05.



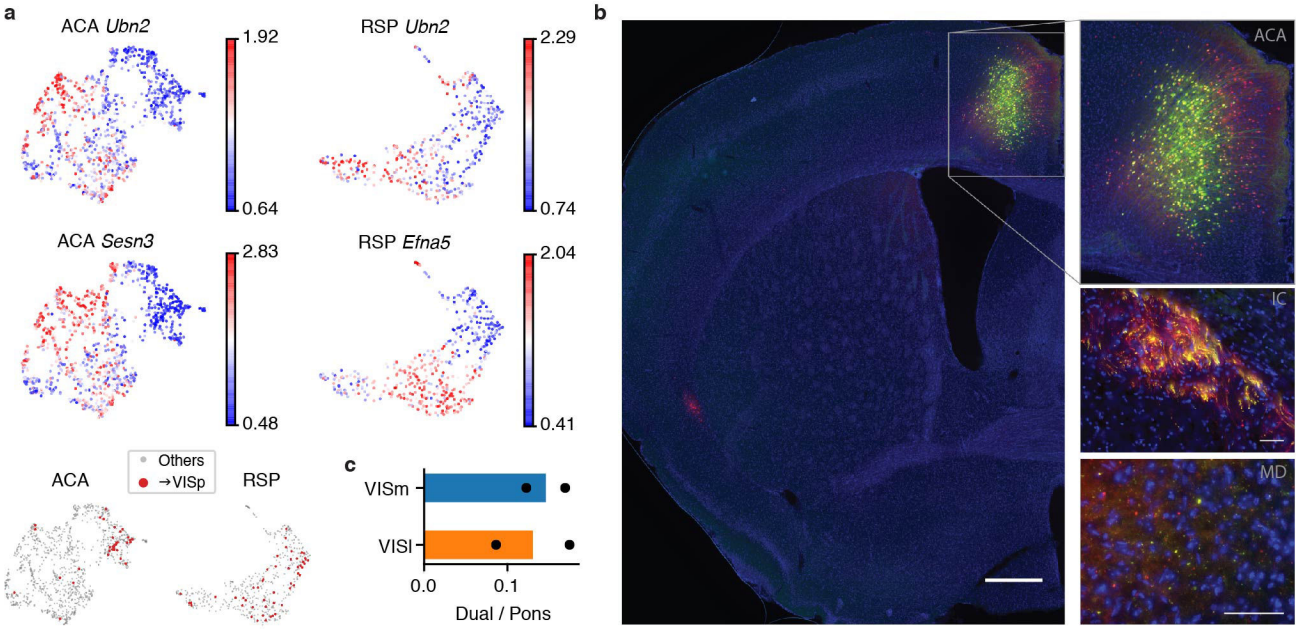
Extended Data Fig. 8. AUROC of ET target pairs within and cross source regions.

AUROC of models trained and tested in the same source region (**a**) or models tested in all source regions after trained in each one of them (**b**) using 100 kb bin mCH as features. Training and testing sets were randomly split. **c,** Overlap score between each pair of targets in each source region. **d,** The proportion of double labeled cells versus the AUROC score to distinguish →SC versus →pons neurons across source areas.



Extended Data Fig. 9. Integration of L5 ET cells from Epi-Retro-Seq and Retro-Seq.

a-c, L5 ET ALM cells in SMART-Seq (n=365) colored by clusters (**a**), major target regions (**b**), and detailed target regions (**c**). Epi-Retro-Seq cells were greyed. **d-i,** L5 ET cells in Epi-Retro-Seq from all source regions (n=4,176) colored by MOp clusters (**d**), SSp clusters (**e**), sources (**f**), targets (**g**), and gene body mCH of *Slco2a1* (**h**) and *Astn2* (**i**).



Extended Data Fig. 10. Validation of L5 ET+CC neurons.

a, UMAP of ACA (n=1,131) and RSP (n=516) L5 ET cells colored by gene body mCH of example genes. *Ubn2* shows hypomethylation in the cluster enriching VISp-projecting neurons in both ACA and RSP, while *Sesn3* and *Efna5* are hypomethylated in the cluster only in ACA or RSP, respectively. VISp-projecting cells are shown in red at the bottom. **b,** By injecting AAVretro-Cre in VISp and AAV-FLEX-GFP in ACA, the axon terminals of ACA→VISp neurons were also observed in internal capsule (IC) and mediodorsal nucleus of thalamus (MD). Scale bars: 500 μm (left) and 50 μm (right in IC and MD). **c,** The proportion of double labeled neurons that project to both VISp and pons, out of neurons projecting to pons in VISm and VISl. n=2 biological replicates are shown as individual points.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Dr. Kai Zhang for advice on the PageRank algorithm, and Dr. Jesse R. Dixon for insightful comments. We are grateful to Dr. Michael Nunn for help with management of the project. This work is supported by NIMH U19MH114831 to E.M.C and J.R.E and by NIMH R01MH063912 and NEI R01EY022577 to E.M.C. M.A.K. is supported by NEI F31 EY028853. The Flow Cytometry Core Facility of the Salk Institute is supported by funding from NIH-NCI CCSG: P30 014195. J.R.E is an investigator of the Howard Hughes Medical Institute.

References

1. Mukamel EA & Ngai J Perspectives on defining cell types in the brain. Curr. Opin. Neurobiol 56, 61–68 (2019). [PubMed: 30530112]
2. Economo MN et al. Distinct descending motor cortex pathways and their roles in movement. Nature 563, 79–84 (2018). [PubMed: 30382200]
3. Chen X et al. High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing. Cell 179, 772–786.e19 (2019). [PubMed: 31626774]
4. Klingler E, Prados J, Kebschull JM, Dayer A & Zador AM Single-cell molecular connectomics of intracortically-projecting neurons. BioRxIV (2018).
5. Kim D-W et al. Multimodal Analysis of Cell Types in a Hypothalamic Node Controlling Social Behavior. Cell 179, 713–728.e17 (2019). [PubMed: 31626771]
6. Tasic B et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature 563, 72–78 (2018). [PubMed: 30382198]
7. Hodge RD et al. Conserved cell types with divergent features in human versus mouse cortex. Nature 573, 61–68 (2019). [PubMed: 31435019]
8. Luo C et al. Robust single-cell DNA methylome profiling with snmC-seq2. Nat. Commun 9, 3824 (2018). [PubMed: 30237449]
9. Tervo DGR et al. A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. Neuron 92, 372–382 (2016). [PubMed: 27720486]
10. Mo A et al. Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron 86, 1369–1384 (2015). [PubMed: 26087164]
11. Luo C et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science 357, 600–604 (2017). [PubMed: 28798132]
12. Luo C et al. Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. bioRxiv 2019.12.11.873398 (2019) doi:10.1101/2019.12.11.873398.
13. Lister R et al. Global epigenomic reconfiguration during mammalian brain development. Science 341, 1237905 (2013). [PubMed: 23828890]

14. Price AJ et al. Divergent neuronal DNA methylation patterns across human cortical development reveal critical periods and a unique role of CpH methylation. Genome Biol 20, 196 (2019). [PubMed: 31554518]
15. Fejtova A et al. Dynein light chain regulates axonal trafficking and synaptic levels of Bassoon. J. Cell Biol 185, 341–355 (2009). [PubMed: 19380881]
16. Sanders SJ et al. Progress in Understanding and Treating SCN2A-Mediated Disorders. Trends Neurosci 41, 442–456 (2018). [PubMed: 29691040]
17. Zingg B et al. Neural networks of the mouse neocortex. Cell 156, 1096–1111 (2014). [PubMed: 24581503]
18. Wilson PM, Fryer RH, Fang Y & Hatten ME Astn2, a novel member of the astrotactin gene family, regulates the trafficking of ASTN1 during glial-guided neuronal migration. J. Neurosci 30, 8529–8540 (2010). [PubMed: 20573900]
19. Lionel AC et al. Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. Hum. Mol. Genet 23, 2752–2768 (2014). [PubMed: 24381304]
20. Harris KD & Shepherd GMG The neocortical circuit: themes and variations. Nat. Neurosci 18, 170–181 (2015). [PubMed: 25622573]
21. Nelson A et al. A circuit for motor cortical modulation of auditory cortical activity. J. Neurosci 33, 14342–14353 (2013). [PubMed: 24005287]
22. Veinante P & Deschênes M Single-cell study of motor cortex projections to the barrel field in rats. J. Comp. Neurol 464, 98–103 (2003). [PubMed: 12866130]
23. Fries W, Keizer K & Kuypers HG Large layer VI cells in macaque striate cortex (Meynert cells) project to both superior colliculus and prestriate visual area V5. Exp. Brain Res 58, 613–616 (1985). [PubMed: 3839191]
24. vogt Weisenhorn DM, Illing RB & Spatz WB Morphology and connections of neurons in area 17 projecting to the extrastriate areas MT and 19DM and to the superior colliculus in the monkey Callithrix jacchus. J. Comp. Neurol 362, 233–255 (1995). [PubMed: 8576436]
25. Wang Q et al. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. Cell 181, 936–953.e20 (2020). [PubMed: 32386544]
26. Lacar B et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. Nat. Commun 7, 11022 (2016). [PubMed: 27090946]
27. Martin M Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12 (2011).
28. Krueger F & Andrews SR Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572 (2011). [PubMed: 21493656]
29. Schultz MD et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523, 212–216 (2015). [PubMed: 26030523]
30. Frankish A et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47, D766–D773 (2019). [PubMed: 30357393]
31. Wolf FA, Angerer P & Theis FJ SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15 (2018). [PubMed: 29409532]
32. Wolock SL, Lopez R & Klein AM Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 8, 281–291.e9 (2019). [PubMed: 30954476]
33. Liu H et al. DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution. bioRxiv 2020.04.30.069377 (2020) doi:10.1101/2020.04.30.069377.
34. Zhang K, Wang M, Zhao Y & Wang W Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development. Sci Adv 5, eaav3262 (2019). [PubMed: 30944857]
35. Khan A et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res 46, D260–D266 (2018). [PubMed: 29140473]
36. Bailey TL, Johnson J, Grant CE & Noble WS The MEME Suite. Nucleic Acids Res 43, W39–49 (2015). [PubMed: 25953851]

37. McLean CY et al. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol 28, 495–501 (2010). [PubMed: 20436461]

38. Hie B, Bryson B & Berger B Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol 37, 685–691 (2019). [PubMed: 31061482]

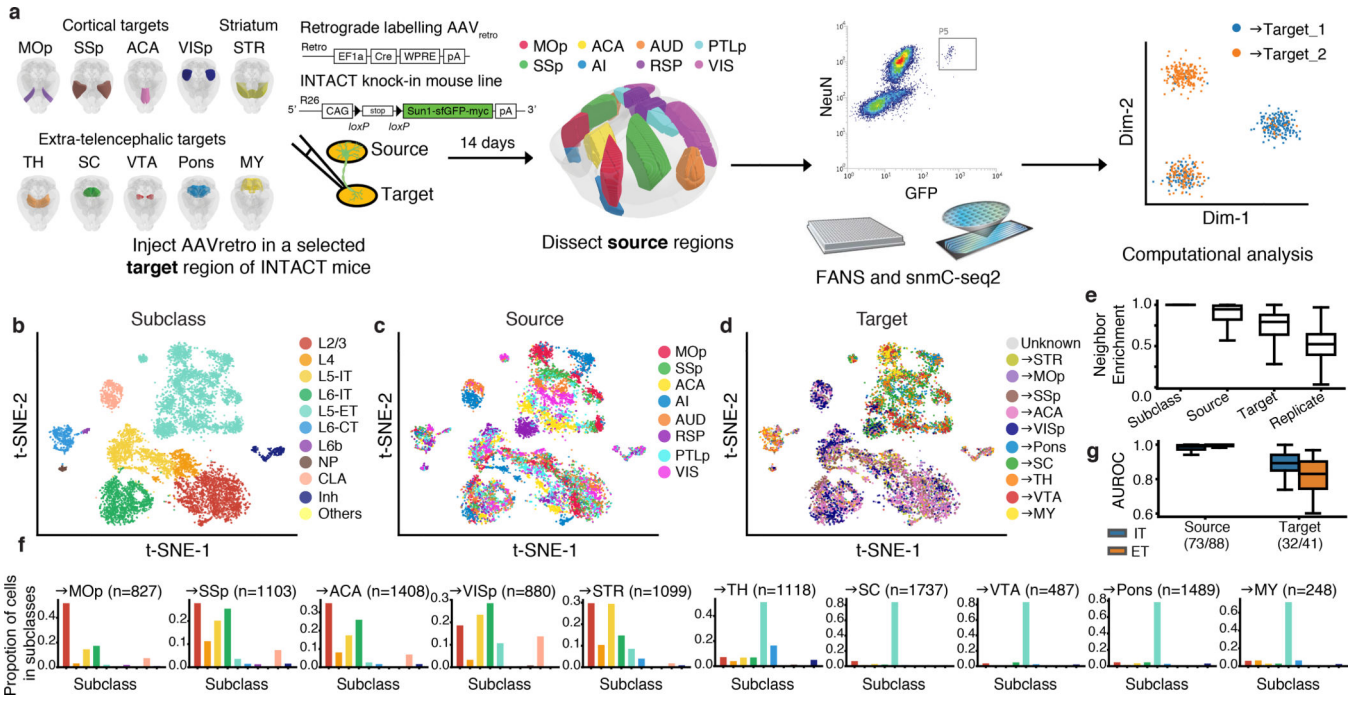


Fig. 1 |. The epigenomic landscape of cortical projection neurons.

a, Schematics of Epi-Retro-Seq workflow. All brain atlas images were created based on © 2017 Allen Institute for Brain Science. Allen Brain Reference Atlas. Available from: atlas.brain-map.org and Wang et al.[25] **b-d,** Two-dimensional t-distributed stochastic neighbor embedding (t-SNE) of 11,827 cortical neuron nuclei based on mCH levels in 100 kb genomic bins, colored by subclass (**b**), the source region of neurons (**c**), or their projection target (**d**). **e,** Neighbor enrichment scores of cells categorized by subclass (n=11,827), source (n=11,827), target (n=10,396), and replicate (n=11,638). **f,** The distribution across cell subclasses of neurons that projected to each IT (left) or ET (right) target. **g,** AUROC of source pairs and target pairs computed for IT and ET neurons based on gene body mCH (n=73/88/32/41). The elements of all boxplots are defined as: center line, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range. NP, near-projecting; Inh, inhibitory.

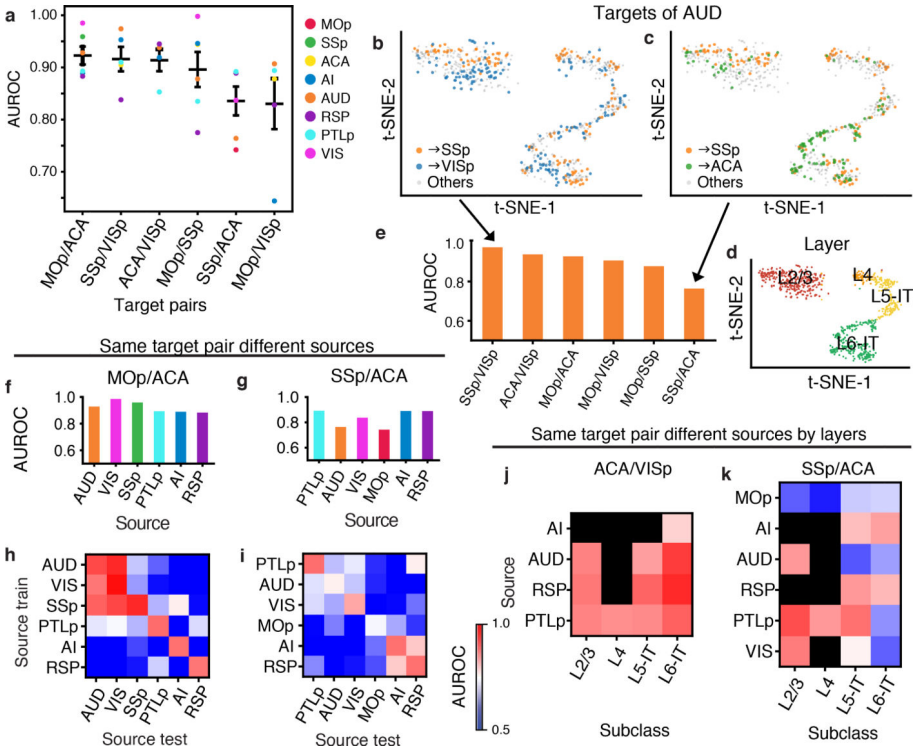


Fig. 2 |. Epigenetic differences between IT neurons projecting to different targets.

a, AUROC to distinguish cortical neurons projecting to one cortical target versus another. Data are presented as mean +/− standard error of the mean (n=6,5,4,6,6,5 sources). **b-d,** t-SNE of AUD neurons in the IT subclasses (n=737) colored by projections **(b, c**) and subclasses (**d**). **e,** AUROC to distinguish AUD neurons projecting to each target pair. **f, g,** The AUROC for comparisons between →MOp versus →ACA neurons from different sources (**f**), and between →SSp versus →ACA neurons from different sources (**g**). **h, i,** Heatmaps of AUROC from prediction models that were trained on one source (row) and tested on another source (column) to distinguish between neurons projecting to →MOp versus →ACA (**h**), or between →SSp versus →ACA neurons (**i**). **j, k,** Heatmaps of AUROC from prediction models that were trained and tested on neurons from each cortical layer (column) in each source(row), to distinguish between →ACA versus →VISp neurons (**j**), or between →SSp versus →ACA neurons (**k**).

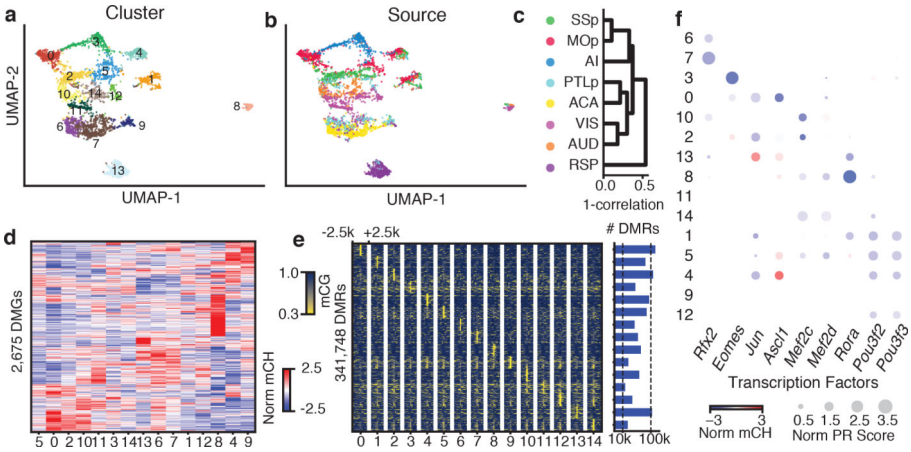


Fig. 3 |. Epigenetic diversity of L5 ET neurons.

a, b, Fifteen clusters of L5 ET neurons (n=4,176) shown on the uniform manifold approximation and projection (UMAP) plot, colored by cluster (**a**), or the source of neurons (**b**). **c,** Dendrogram shows the correlations between mCH profiles of L5 ET neurons from different sources. **d,** Gene body mCH levels in each cluster of 2,675 CH-DMGs that were identified in pairwise comparisons between L5 ET clusters. **e,** 341,748 differentially methylated regions (CG-DMRs) were identified across the 15 L5 ET clusters. The mCG levels at CG-DMRs and their 2.5kb flanking genomic regions in each cluster were visualized in the heatmap (left). The numbers of CG-DMRs hypo-methylated in each cluster were plotted in the bar chart (right). **f,** Examples of some predicted key regulator TFs. The size of each dot represents the normalized PageRank score of the TF. The color of the dot represents the gene body mCH of the TF in the corresponding L5 ET cluster.

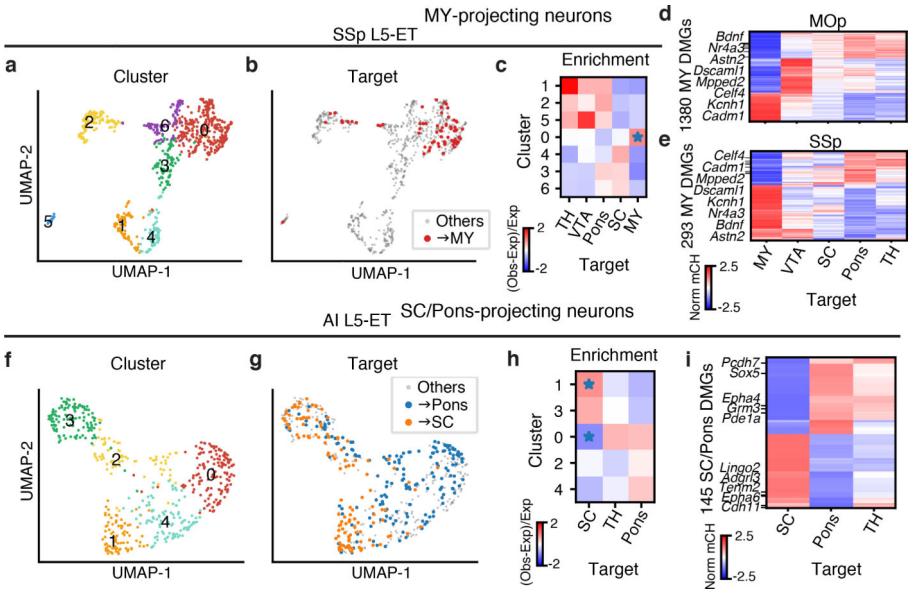


Fig. 4 |. Epigenetic differences between L5 ET neurons projecting to different targets.

UMAP of SSp (**a**, n=884) or AI (**f**, n=531) L5 ET neurons by 100 kb-bin mCH are colored by clusters (**a, f**) or projection targets (**b, g**). The enrichment of SSp (**c**) or AI (**h**) neurons projecting to each target in each cluster (* represents FDR<0.05). Gene body mCH levels of the CH-DMGs in MOp (**d**) or SSp (**e**) between neurons projecting to MY and other ET targets, or in AI between SC and Pons projecting neurons (**i**). Values are Z-score normalized by rows. Examples of CH-DMGs hypo-methylated in both MOp→MY and SSp→MY neurons are labeled in d and e.

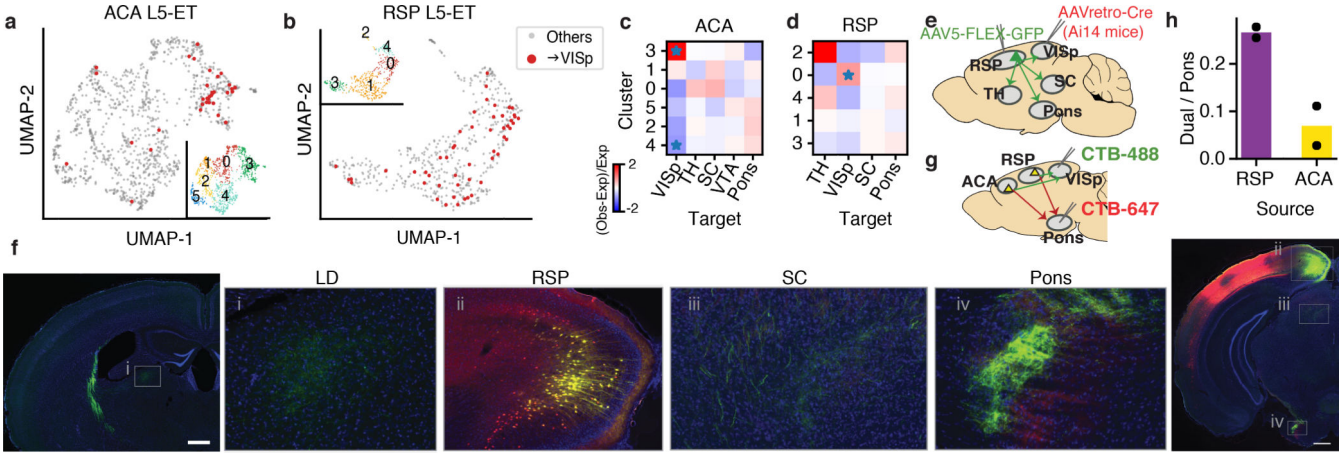


Fig. 5 |. A L5 ET neuron type that projects to both ET and cortical targets (L5 ET+CC).

a, b, UMAP embedding of ACA (**a**) or RSP (**b**) L5 ET neurons (n=1,131/516) using mCH in 100 kb bins, colored by projection targets (ACA or RSP→VISp in red, n=36/51) or clusters (Inset). **c, d,** ACA→VISp neurons were enriched in ACA L5 ET cluster 3 and depleted from cluster 4 (**c**). RSP→VISp neurons were enriched in RSP L5 ET cluster 0 (**d**). (* indicating FDR<0.05). **e,** Illustration of the anatomical experiment to validate the existence of L5 ET+CC cell type. **f,** VISp neurons at the AAVretro-Cre injection site were labeled by tdTomato (red). RSP→VISp neurons were labeled with GFP (green), among which RSP→VISp neurons at the AAV5-FLEX-GFP injection site were labeled with both tdTomato and GFP (yellow; inset ii). Scale bars: 500 μm (low magnification). **g,** Illustration of dual retrograde tracer injections into pons and VISp. **h**, Proportion of double-labeled neurons (projecting to both pons and VISp) among all pons-projecting neurons in different sources. n=2 biological replicates are shown as individual points.